



**HAL**  
open science

# Ecological niche modelling and its application to environmentally acquired diseases, the case of *Mycobacterium ulcerans* and the Buruli ulcer

Kevin Carolan

► **To cite this version:**

Kevin Carolan. Ecological niche modelling and its application to environmentally acquired diseases, the case of *Mycobacterium ulcerans* and the Buruli ulcer. Infectious diseases. Université Montpellier II - Sciences et Techniques du Languedoc, 2014. English. NNT : 2014MON20178 . tel-01818034

**HAL Id: tel-01818034**

**<https://theses.hal.science/tel-01818034>**

Submitted on 18 Jun 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Université Montpellier 2

Ecole doctorale SIBAGHE

Kevin Carolan

Ecological niche modelling and its application to environmentally acquired diseases, the case of *Mycobacterium ulcerans* and the Buruli ulcer

Co-encadrants: Danny Lo Seen & Jean-François Guégan

Thèse pour obtenir le grade de DOCTEUR DE L'UNIVERSITÉ  
MONTPELLIER II

Présenté et soutenu publiquement le 12 décembre 2014



## Jury

Bernard Cazelles	Professeur Ecole Normale Supérieure	Rapporteur
Simon Hay	Professor at Oxford University	Rapporteur, Examineur
Laurent Gavotte	Maitre de conférences, University de Montpellier II	Examineur
Jean-François Guégan	Directeur de recherche à l'IRD, professeur à l'EHESP	Co-directeur de thèse
Danny Lo Seen	Chercheur à La Centre de coopération internationale en recherche agronomique pour le développement	Co-directeur de thèse
Sébastien Gourbiere	Chercheur à Université de Perpignan	Examineur

## Contents

<b>Chapter 1 –Introduction</b>	1
What is the ecological niche?	2
Why is knowledge of the ecological niche useful?	3
Tools to represent the ecological niche,	3
Not just the condition, but also the spatial configuration of habitat influences suitability.	4
Examples of techniques for modelling the ecological niche.	7
Limitations of ecological niche modelling	12
Moving up and down scales; the niche model changes from one scale to another	12
Worldwide prevalence and pathology of Buruli ulcer	15
International response to Buruli ulcer emergence	15
Evolution of <i>M. ulcerans</i>	18
Environmental distribution of <i>M. ulcerans</i>	20
Routes by which <i>M. ulcerans</i> infects humans	22
Overview of the thesis	24
References	25
<b>Chapter 2: How many samples do we need to describe the spatial distribution of an emerging pathogen?</b>	30
Foreword	31
Abstract	32
Introduction	32
Materials and Methods	37
Results	42
Discussion	44
References	46
<b>Chapter 3: Topography and land cover of watersheds predicts the distribution of the environmental pathogen <i>Mycobacterium ulcerans</i>.</b>	48
Foreword	49
Author summary	50
Background	50
Methodology	53
Results	69
Discussion	73
Conclusion	75
References	76
<b>Chapter 4: Ecological niche modelling of Hemipteran insects in Cameroon; the paradox of a vector-borne transmission for <i>Mycobacterium ulcerans</i>, the causative agent of Buruli ulcer.</b>	80
Foreword	81
Abstract	83
Background	84
Materials and Methods	86
Results	92
Discussion	97
References	100

<b>Chapter 5: Synthesis.</b>	102
Introduction	103
Ecological niche models; sample size, surface area and performance	103
The ecological niche of <i>M. ulcerans</i>	104
The relationship between Belostomatidae, Naucoridae and Buruli ulcer	109
Conclusion, practical applications, and questions for future research	112
Conclusion	115
References	116

<b>Chapter 6: Appendix.</b>	117
Foreword	118

Marion *et al.*, Buruli ulcer in south western Nigeria Emerging infectious diseases

García-Peña *et al.*, Niche-based extinction of species boosts presence of the environmentally-acquired pathogen, *Mycobacterium ulcerans*

Landier *et al.*, Space-time dynamics and landscape-associated risk of Buruli ulcer in Akonolinga, Cameroon

## List of Figures

<i>Figure</i>	<i>Page</i>
Figure 1	5
Figure 2	6
Figure 3	8
Figure 4	9
Figure 5	10
Figure 6	14
Figure 7	14
Figure 8	15
Figure 9	17
Figure 10	19
Figure 11	21
Figure 12	22
Figure 13	23
Figure 14	31
Figure 15	34
Figure 16	35
Figure 17	36
Figure 18	38
Figure 19	40
Figure 20	42
Figure 21	43
Figure 22	54
Figure 23	59
Figure 24	64
Figure 25	67
Figure 26	68
Figure 27	70
Figure 28	72
Figure 29	82
Figure 30	86
Figure 31	89
Figure 32	91
Figure 33	92
Figure 34	95
Figure 35	96
Figure 36	113

## List of Tables

<i>Table</i>	<i>Page</i>
Table 1	55
Table 2	56
Table 3	61
Table 4	62
Table 5	64
Table 6	65
Table 7	71
Table 8	88
Table 9	93
Table 10	95

## List of Equations

<i>Equation</i>	<i>Page</i>
Equation 1	57
Equation 2	87

## Content summary

33,112 Words  
118 Pages  
6 Chapters  
36 Figures  
10 Tables  
2 Equations

## Acknowledgements

Grateful thanks go of course to my supervisors, Danny Lo Seen and Jean-François Guégan. The last three years have been difficult at times, but their consistent support, advice and input has been invaluable. Danny and Jean-François have been two of the best advisors anyone could have asked for.

I would also like to thank my thesis committee, Annelise Tran, Benjamin Roche and Serge Morand for invaluable insight and advice during my comité de these, which proved a key turning point in my research. Simon Hay and Bernard Cazelles also deserve special thanks for undertaking the task of rappourter (examiner) of this thesis. In the past 3 years I have been supported at the Institut de Recherche pour le Développement and Maison de la Télédétection, and my thanks go to the staff at IRD and MTD, and the Institute Pasteur, LabEx CEBA for additional support.

The co-authors of my papers, Andres Garchitorena, Daniel Sanhueza, Gabriel Garcia-Pena, Jordi Landier, Estelle Marion, Aaron Morris, Solange Meyin À Ebong, Gaëtan Texier, Laurent Marsollier, Philippe Le Gall, Arnaud Fontanet, Rodolphe Gozlan and Sara Eyangoh, provided excellent support and comments on various revisions of my papers.

Thank you to all the PhD's at IRD, Manon, Romain, Marlene, Elsa and Elsa, the Aussie contingent; Mitch, Elsa, Tim and Nicole, the memes *et al*; Charlotte, Viola, Christian, Francisco, Rachel, Alex, Eva, Orrin, Scott, Miles, Georgia, Chris, and Katie, for support, comfort and burgers.

To Lidia: the most awesome person I know, who kept me sane, made me smile, and fallen in love for the last year and a half. And finally to my Mum, my Dad and my family; for being themselves.



## Chapter 1: Introduction

## What is the ecological niche?

Every species in nature has a range of environments within which it can survive and maintain a population. Outside this range the conditions are unsuitable, for example they may be too hot or too cold, too wet or too dry, too many predators or not enough prey, or limiting in some other way. The set of suitable habitat is termed the ecological niche of the species (Hutchinson 1957, Hirzel *et al.* 2002, Soberon and Nakamura 2009, Elith *et al.* 2010).

Species compete with each other for particular resources, for example food and water. Within their ecological niche they are better adapted to compete for a particular resource, and other species are excluded from the resource. For example, cacti are adapted to dry arid environments, and will outcompete other plants in these regions. The niches of two species may be similar, for example two species may have similar tolerance to temperature (most Mediterranean plants can survive in similar climates) but there is normally some differentiating factor, such as soil *pH* or disease resistance, that differentiates the two niches, resulting in a unique niche for every species. This is the principle of competitive exclusion. If the niches overlap there is the possibility that species will compete for resources. If competition occurs, one or both species will suffer, and if these niches overlap completely (if they compete for exactly the same resources) one species will be driven to extinction. For this reason we expect that every species has a unique ecological niche, a set of habitat suitable for that species, that describes the environment where we are likely to find the species.

The concept of the ecological niche has changed through time. In 1917 Joseph Grinnell observed that the distribution of the California Thrasher (*Toxostoma redivivum*) was accurately described by vegetation, topography and climate range of California (Grinnell 1917). Grinnell conceived of the niche as an entity in space, and in 1929 Charles Elton expanded this to focus on the importance of interactions between species. Elton viewed the niche as a property of the interaction between species, describing the niche as a property of the biotic community rather than the particular species. The next major change came from George Hutchinson, who expounded upon the concepts of the fundamental and realised ecological niche. For Hutchinson, the fundamental niche was the range of conditions within which a species could survive on the basis of its physiology, while only a subset of this is realised (the realised niche) as the species will not encounter all these conditions in nature (Hutchinson 1957). The concept of the ecological niche continues to evolve, and the practical applications of this ecological theory continue to expand.

## Why is knowledge of the ecological niche useful?

We often wish to infer the distribution of this suitable habitat (the ecological niche) for a species of interest. Understanding the distribution of this habitat has a broad range of applications, from the conservation of endangered species (Peterson *et al.* 2003), to the control of a harmful species (Tran *et al.* 2008, Bhatt *et al.* 2013, Gething *et al.* 2010, Simard *et al.* 2009), which in the case of pathogens or their vectors can enable the control of disease. Understanding the distribution of suitable habitat of a pathogen enables control, avoidance or other management opportunities for a disease (Hay *et al.* 2013). The goal of this thesis is to describe the distribution of suitable habitat for a particular pathogen, *Mycobacterium ulcerans*, and the insects suspected to be its vectors, to facilitate control of the Buruli ulcer disease.

## Tools to represent the ecological niche

When we describe the niche, we can consider it as a hypervolume (Hutchinson 1957). For example, in Figure 1 we have represented the niche of a hypothetical species with three variables, *pH*, UV light intensity and oxygen concentration. We can include further variables if needed; temperature, predation pressure, parasitism, population density, depending on the particular variables we suspect are most important for describing the distribution of that species and available in the area of study. We describe this as an *n*-dimensional hypervolume (Hutchinson 1957). Here, *n* is the number of parameters being described, and each parameter (*pH*, oxygen, UV light) is an axis on the volume. If we describe one dimension, we have a simple line graph, two dimensions are a square, three are a cube, four are a tesseract and so on. The principle of competitive exclusion still applies; the extent to which the two species overlap in the hypervolume describe the potential for competition in the hypervolume. A number of tools have been developed to represent this hypervolume, which can be particularly useful both for the purpose of communication and as a tool to help clarify research goals.

The concept of the niche as a hypervolume is a useful tool that was developed to facilitate thinking about the niche. While it is useful, more than three parameters is difficult to represent, so we often simplify the hypervolume to a Venn diagram (Soberon and Nakamura 2009), as shown in Figure 1. The use of a Venn diagram to describe the niche is becoming common, particularly in the format of the Biotic-Abiotic-Migratory diagram (BAM diagram), by Soberon and Nakamura 2009. The biotic-abiotic-movement diagram was originally proposed to represent the distribution of suitable habitat using three parameters, the combination of suitable biotic conditions (predators, prey, parasites), suitable abiotic conditions (temperature, precipitation), and the environment accessible to a species,

described as movement. This is one particular formation, the flexibility of the BAM format means we can easily use the same principle to describe the distribution of just abiotic variables, or of multiple biotic variables. The movement section of the BAM diagram also highlights the importance of accessibility, discussed below.

### **Not just the condition, but also the spatial configuration of habitat influences suitability**

When the niche hypervolume is projected into real space it describes the distribution of suitable and unsuitable habitats for a species. This is the basis of the fundamental (hypervolume, driven by the species) and realised ecological niche (spatial, driven by the environment), as discussed by Hutchinson. However, the resulting spatial configuration of this suitable habitat is important for the persistence of the species. Habitat may be suitable, but if the physical area of the habitat is too small or does not persist for a sufficient time, the habitat will not support a large population of the species, as in Figure 2. Habitat that is patchy or fragmented leads to higher local risk of extinction, and meta-community dynamics become important in maintaining the species (Figure 2 region 3, Gravel *et al.* 2006). If the species can get reintroduced often enough it can persist in an area, in the face of local extinction (Lockwood *et al.* 2005).

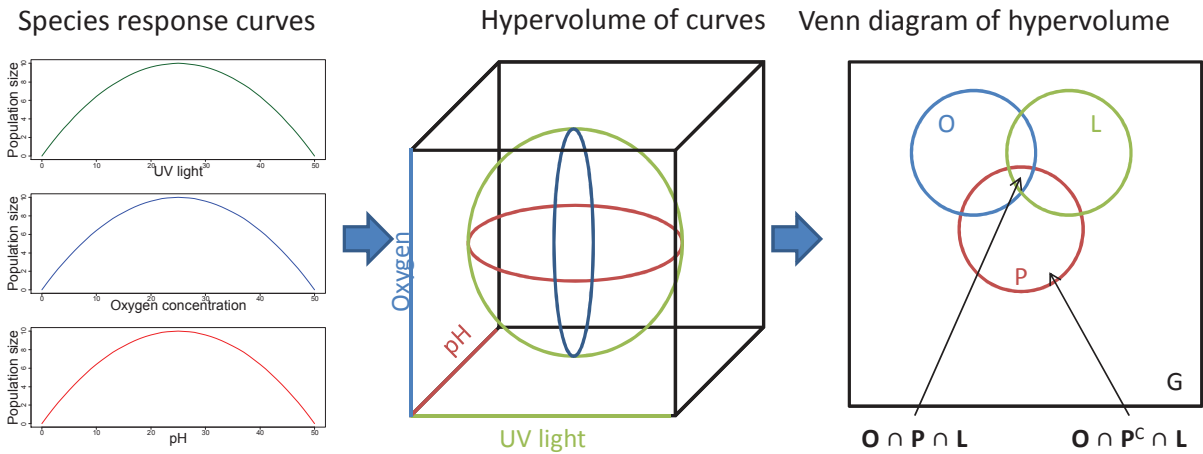


Figure 1. The ecological niche as a hypervolume. Here, we describe how a given species responds to three variables (this is a hypothetical species, to illustrate the example). Three variables are used, intensity of ultraviolet light (in green), oxygen concentration (in blue) and *pH* (in red). There is a certain range of *pH* that is suitable for the species, and so for oxygen and UV light. We can represent this as a three dimensional cube or, if there are more than three dimensions, a hypervolume. Each of the variables important in the species distribution is an axis in the hypervolume. In this example, this results in a sphere of suitable habitat within the hypervolume. Complex hypervolumes can be represented as a Venn diagram. In the Venn diagram, the terminology  $\cap$  is read as 'intersection', meaning the area of overlap between two regions. The term  $^c$  refers to 'complement', the area outside the region, so  $P^c$  is the area of habitat unsuitable due to unsuitable *pH* values. The area  $O \cap P \cap L$  is the combination of suitable environmental parameters. The term  $G$  refers to the entire region studied, the extent volume of the hypervolume. As discussed in Chapter 2, appropriately setting this extent is important in model performance. An ecological niche model of aquatic insects from the Families Belostomadiae and Naucoridae (the suspected vectors of our subject species, *M. ulcerans*) is constructed in Chapter 4, based on the principles developed.

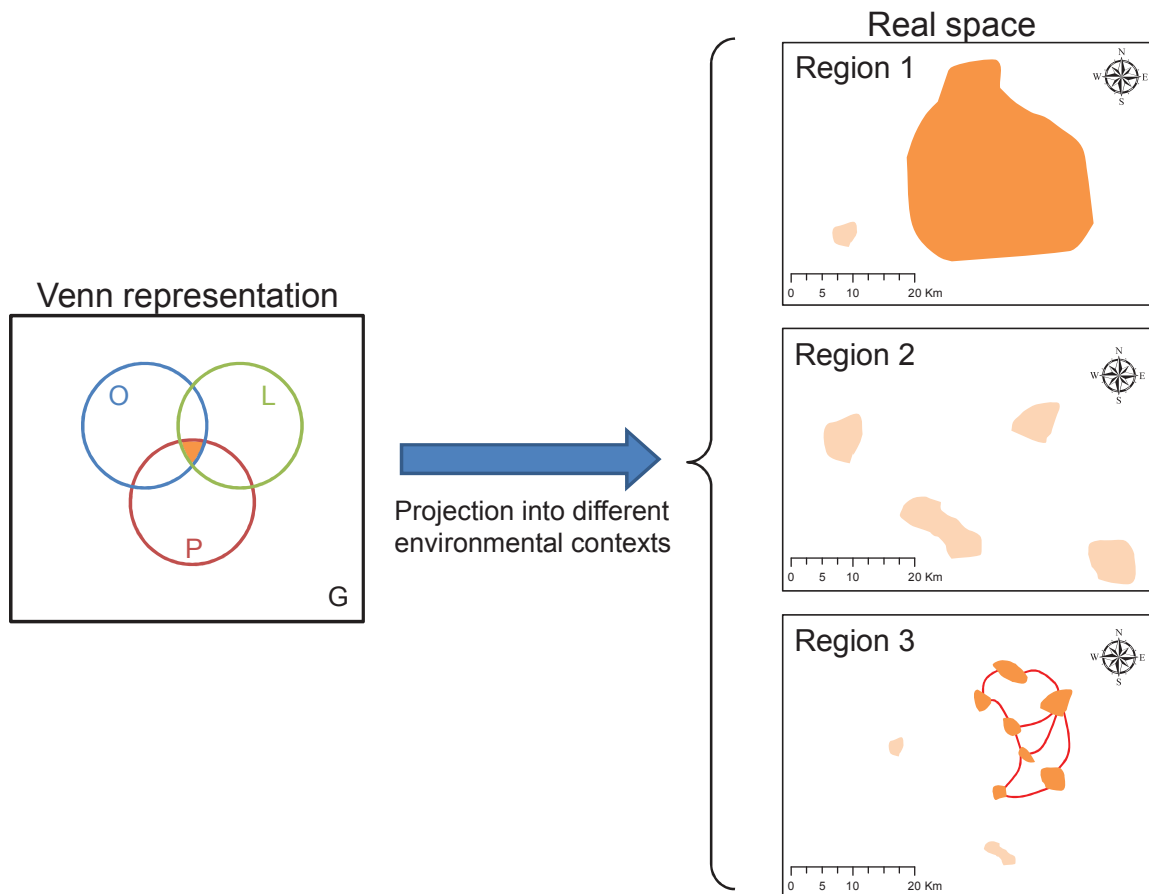


Figure 2. The importance spatial distribution of suitable habitat for a species. The ecological niche, here represented as a Venn diagram (as in Figure 1) is projected from ecological space into real space. The area of where suitable conditions co-occur ( $O \cap P \cap L$ ) is the orange area, which is where we expect the species to occur (Region 1). However, certain areas are physically too small to maintain a viable population (light orange) and are uninhabited, despite being suitable for the species. Because the environmental variables can change from region to region, the spatial assemblage of suitable habitat will also vary, even if the ecological niche is the same. This is represented in the difference between Region 1 and Region 2, where the spatial distribution of oxygen, UV light and  $pH$  change, such that all habitat is too small. This can also be temporal in nature, rather than between areas; for example,  $pH$  may be suitable for a short time in a particular season, but this time may be insufficient to permit colonisation by the species. This can be particularly important for slow growing species. Meta-community dynamics become important at this point, as shown in Region 3. Here, the suitable areas are spatially separate and individually are not able to maintain a population, but because of connections to other areas (red lines) the species is able to maintain itself. This can be important for many highly mobile species. Practically, this means local control of a population of a species may be achieved by non-local interventions. An example of this is found in the conservation procedures proposed for endangered Ethiopian wolves, where vaccination strategies are aimed at a different species, wild dogs, to control the movement of the rabies virus. In that scenario, the wolves and dogs can be thought of different ‘patches’ of habitat for the virus (Haydon *et al.* 2002).

## Examples of techniques for modelling the ecological niche

When we decide to describe the distribution of suitable habitat, there are a variety of different methods available for modelling the ecological niche of species. The performances of these methods have been compared in a variety of papers (Segurado and Araùjo 2004, Austin 2007, Soberon and Nakamura 2009). The purpose section is not to extensively compare each modelling method, but to give a brief overview of different methods available in terms of their assumptions and data requirements. In principle we should try to understand how the structure of our data and the goal of our study meet the assumptions of different techniques, and decide which method to use on this basis.

The first division in techniques is between correlative and mechanistic niche models. Correlative niche models attempt to correlate the species population size to the observed environmental values (Figure 3), while mechanistic models are based on knowledge of the physiology of the species (Figure 4). An example of mechanistic niche modelling is found in Kearny *et al.* (2004), who use an understanding of the physiological tolerance of a lizard to temperature, and other environmental factors, to identify the distribution of the species in Australia in space and time. Neither correlative nor mechanistic methods are without limitations, however in this thesis I have used correlative niche techniques as we generally do not have the prerequisite physiological data for our target species, *Mycobacterium ulcerans*.

There are three types of species distribution data used in correlative ecological niche modelling. Presence only data, presence-absence data, and abundance data. These data are different in the questions they can answer and the uses to which they can be applied. Presence only data are easier to collect, but are influenced by sample site bias – choice of sample sites has a large effect on the model (Elidith *et al.* 2010, Chapter 2 of this thesis). On the other hand, absence data is often difficult to interpret; is the species absent, or did we fail to detect it? Some sites are easier to reach and so more intensively or reliably sampled, certain species are easier to detect. For example, when detecting environmental bacteria we often use the polymerase chain reaction (PCR) to detect their DNA. This raises issues of reliability; in certain conditions DNA is easier to detect, causing detectability to vary in space and time, as in Figure 5. This is the case for our subject species, *M. ulcerans*, which is detected environmentally using quantitative PCR. For these reasons presence-absence data generally require more effort to gather. Conclusively stating that a species is absent from a region is difficult. The amount of data, of any type, needed to construct a model varies, and is discussed in detail in Chapter 2.

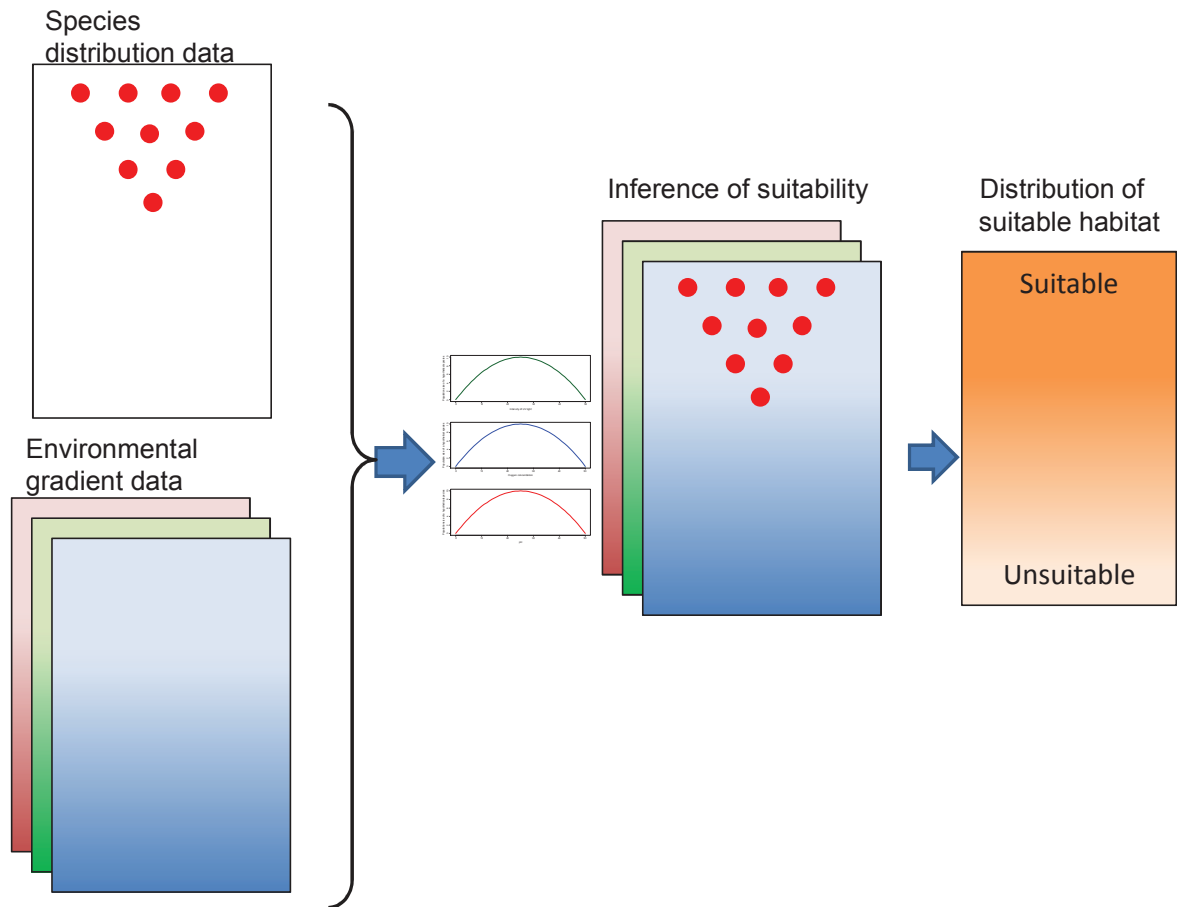


Figure 3. Schematic explanation of correlative niche modelling. Having identified the distribution of the species in space (the red dots), we select certain environmental parameters (gradient data) and identify those parameters that are strongly correlated to the presence of the species. This allows us to describe the distribution of the species beyond the sample sites, but (as is described in Chapter 2) is highly sensitive to the study area, sampling regime, and sample sites selected.



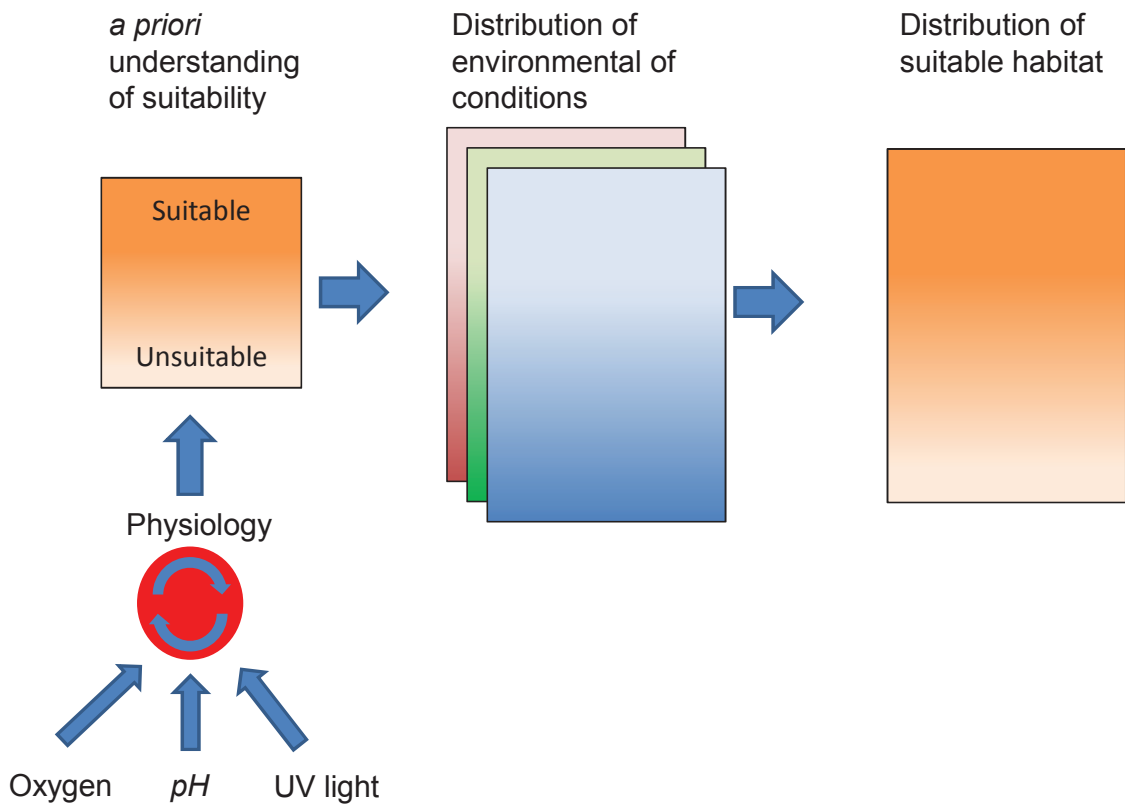


Figure 4. Schematic explanation of mechanistic niche modelling. By studying the physiology of a species, and its responses to different environmental parameters such as oxygen,  $pH$  and UV light we can identify how each parameter is limiting in different ways. This gives us an *a priori* understanding of the niche and what constitutes suitable habitat. We then identify the repartition of these parameters in the environment to describe the expected distribution of the species. Collection of this physiological data often requires intensive laboratory studies, and is impractical for certain species. For example, bacterial species that are difficult to culture make such studies time consuming and difficult.

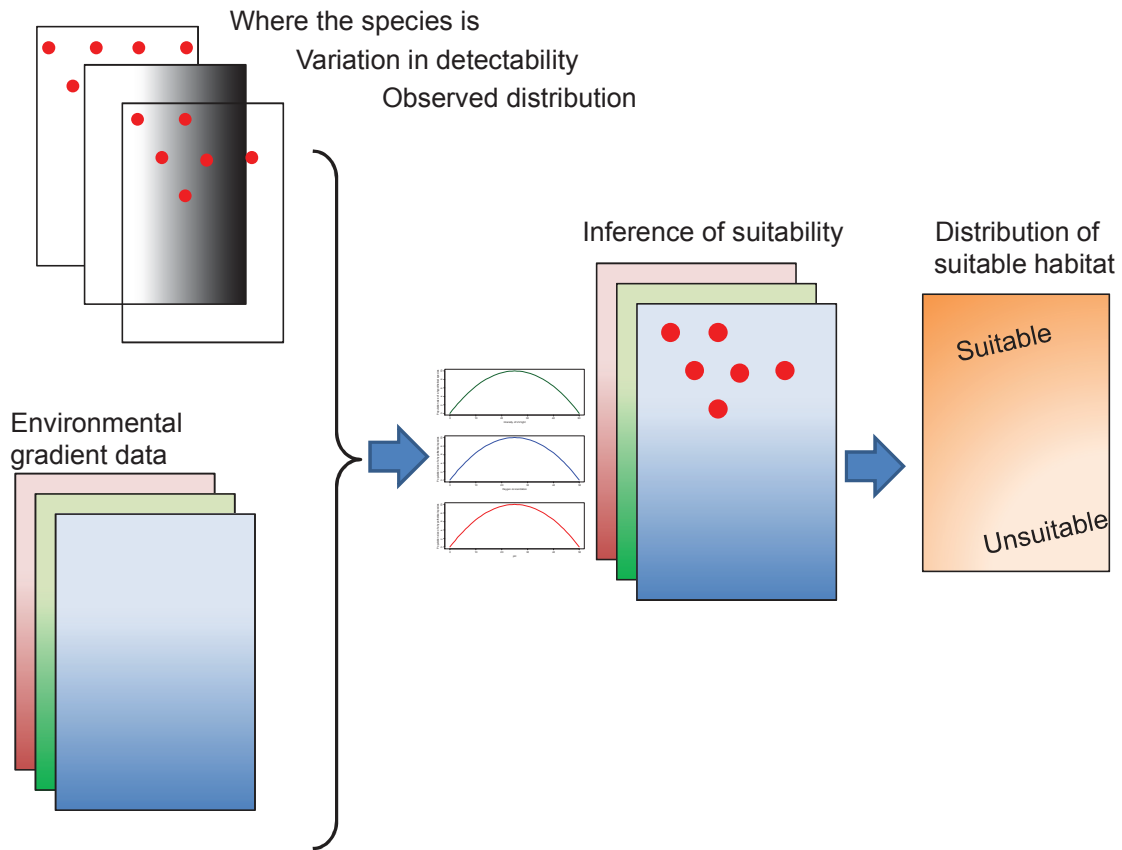


Figure 5. The effect of detectability as a limitation in niche modelling. Here, as discussed in Figure 3, correlative niche modelling uses species distribution data to identify the environment where the species is common. We survey the species and attempt to detect its presence or absence in certain regions, however, the gradient in the black panel (upper left) represents changes in detectability of the species from one region to another. If our ability to detect the species decreases, we falsely identify absence, and receive a confused picture of habitat suitability. Detectability can change for a number of reasons. Certain species are much easier to detect than others, being large and obvious, such as large mammals. Others, such as bacteria (our study species, *M. ulcerans*), are harder to detect. Detectability also changes from one environment to another for the same species. This highlights the importance of consistent detection methodology between different studies. If sampling effort is not consistent the sampling regime can bias our results, giving different impressions of the ecological drives of the species.

Once we have the environmental variables and the species distribution data, a variety of methods are available to describe the correlation between the two. One of the oldest types of model (Guisan *et al.* 2002), linear models fit a response variable  $Y$  (the species distribution data) to an explanatory variable  $X$  (environmental data) such that  $Y=c+mX+e$ , where  $c$  is the intercept,  $m$  is the slope and  $e$  is error. These models assume constant variance, normal error distribution and error independence. Linear regression is a simple method, and makes the basis of many other methods. For example, Glass *et al.* (1995) used logistic regression to model the risk of Lyme disease in Baltimore, finding that particular soils, forests and urban areas were at greater or lesser risk for human infection.

Generalized linear models (GLM) are much more flexible extensions of linear models (Guisan *et al.* 2002). They do not require data to be normally distributed, and can use data with binomial, Poisson or Gamma distributions. With these distributions we can model continuous (eg temperature), categorical (eg land cover) and proportional data. This greater flexibility allows us to describe more complicated relationships between the species distribution and the environment. Generalized additive models (GAM) work in a similar way to GLMs, but they can include nonparametric predictors using an automated smoothing function (Zuur *et al.* 2009). GAM is more flexible than GLM, but has a greater risk of over-fitting. GLM's and GAM's have been used commonly in different applications for ecological niche modelling, for example, Meentemeyer *et al.* (2008) used a GLM to describe the distribution of suitable habitat for the arboreal pathogen *Phytophthora ramorum* in California (Meentemeyer *et al.* 2008).

Another potential method to use in ecological niche modelling is maximum entropy modelling. This method has been increasing in popularity, due to its ability (unlike all the above methods) to use only presence data as opposed to presence-absence or count data. Maximum entropy fits the response variable (species presence data) to the predictor (environment data) assuming the species has the most ecologically efficient distribution. This assumption is not always realistic, for example breeding populations are often in high population densities for a short period of time. Practically, meeting this assumption means that the Kullback–Leibler divergence (a measurement of the difference between two probability distributions) between the curves is minimised (Phillips *et al.* 2008). To avoid overfitting, a particular parameter,  $\tau$ , must be specified.  $\tau$  is based on the prevalence of the species in the study area, and is normally unknown. The method has two notable limitations,  $\tau$  is assumed (from a study of 226 species, Elith *et al.* 2006) and may change, and the area of the study dramatically affects the resulting performance of the model. In other ways maximum entropy is very robust, producing accurate predictions for small datasets (Elith *et al.* 2006). Maximum

entropy, and all of the above methods, are sensitive to the scale the study is conducted at, discussed below.

The above list of modelling methods is not exhaustive, many other methods for niche modelling exist (for example, Ecological niche factor analysis, Hirzel *et al.* (2002), or Genetic algorithm for rule-set prediction, Townsend Peterson, Ball and Cohoon (2002). The purpose of the above section is to introduce some of the basic concepts, to set the rest of the thesis in context. Having selected a method to model the niche, we then need to evaluate the resulting model(s). There are multiple different ways to assess the fit of a model. No one method of model evaluation is best, and each informs us on different aspects of performance. Potential methods for testing models are explored in detail in chapters 2, 3 and 4 of this thesis. Finally, it should be remembered that a well-fitting model may be completely biologically irrelevant: goodness of fit does not mean biological realism, and we can never get away from the need for expert opinion.

### **Limitations of ecological niche modelling**

The basic principle of correlative niche modelling, as discussed above, is to identify suitable habitat by inferring that large populations are present in suitable habitat. If we observe a gradient in population along with an environmental gradient, we assume the habitat with the large population is more suitable. This has certain limitations. We can underestimate or overestimate the size of the population for a number of reasons. The first of these issues is detectability, as discussed above and in Figure 5. The second issue is accessibility of the site by the species. Habitat that we would normally expect to be suitable is often uninhabited by a species due to difficulty in access. This is a large issue in correlative niche modelling, as in practice it is often extremely difficult to differentiate habitat that is unsuitable or simply uninhabited. This has important implications for our interpretation of the resulting models; are we trying to identify habitat suitable for the species to maintain a population (which is perhaps a target of intervention for pathogens, or conservation for endangered species), or a region accessible but not necessarily suitable for a species? Evidence of absence is also difficult to interpret, as discussed above.

### **Moving up and down scales; the niche model changes from one scale to another**

As mentioned above, the scale of the study is important. Scale influences the environmental drivers important in the ecological niche, and the performance of any given model is likely to vary from one scale to another (Tilman *et al.* 1997). At small scales (Figure 6) the suitability will be driven by the physiology of the species. For example, at the local scale *Mycobacterium ulcerans* is influenced by

the oxygen concentration, the intensity of UV light and the *pH* of the environment (Palomino *et al.* 1998, Stinear *et al.* 2007, Merritt *et al.* 2010, Garchitorena *et al.* 2014). While these parameters, and others, will directly affect the growth and survival rate of the bacterium, the distribution of these parameters will in turn be influenced by the local environment. Environmental gradients change rapidly and at short distances (See for example, Caillon *et al.* 2014 and Figure 7). Conditions will change rapidly within and outwith cells (Coutanceau *et al.* 2005), and within outwith biotic hosts. However, in turn, at a landscape scale, the distribution of biotic hosts changes, water *pH* will be influenced by land cover (swamps will have different average *pH* to forests), and the parameters that are most useful at describing the distribution of the bacterium change (Figure 6). As the focus of the inquiry changes (from understanding the physiology of the bacterium to understanding why it is more common in one region of a country than another) the scale of our study, and the resulting ecological niche model, changes.

Herein, I focus on the emerging neglected pathogen *Mycobacterium ulcerans*, and its suspected insect vectors. Using correlative niche modelling, I describe the distribution of environmental variables at the landscape scale associated with the distribution of the pathogen in Cameroon, then tested in French Guiana. Knowledge of the environmental drivers of the spatial distribution of *M. ulcerans* will aid in control of Buruli ulcer disease in endemic countries. A detailed description of the pathogen and the disease, and the expected impact of the thesis, are described below.

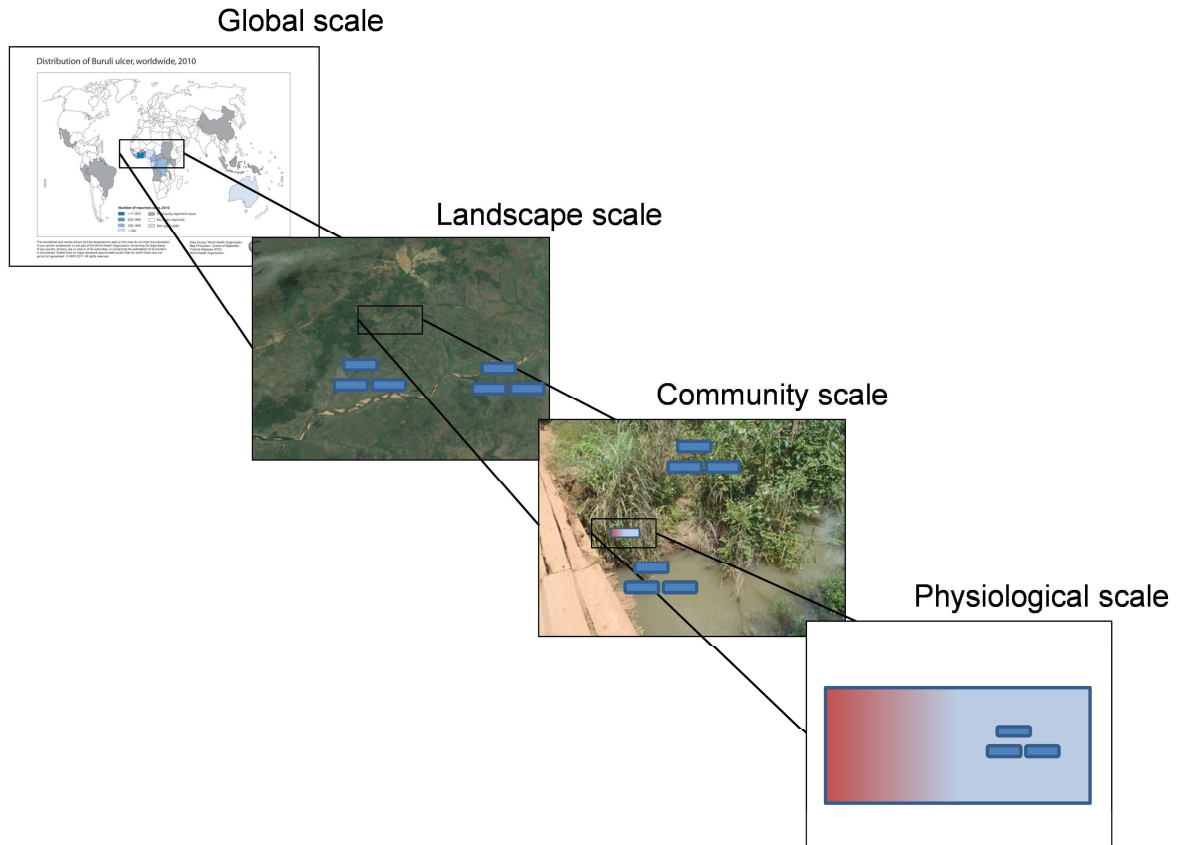


Figure 6. The importance of spatial scale of the study in influencing the niche model. Here, for example, we consider how a bacterium can be driven by different environmental parameters at different scales. Here the blue rectangles represent the presence of the bacterium. At the finest scale, the species is driven by  $pH$  physiology, for example gradients in  $pH$ . At the local or community scale the species, though the species is still influenced by  $pH$ , the distribution of the species is more easily described by those factors that describe the distribution of  $pH$ , for example soil types, plant community, or water quality. The same principle is applied at the larger landscape scale, where the distribution of soil conditions, water quality and plants are described by the shape of the landscape and climate factors such as rainfall. Community scale image © Kevin Carolan, Landscape image reproduced from Google Earth, Global scale image reproduced from (WHO 2010).

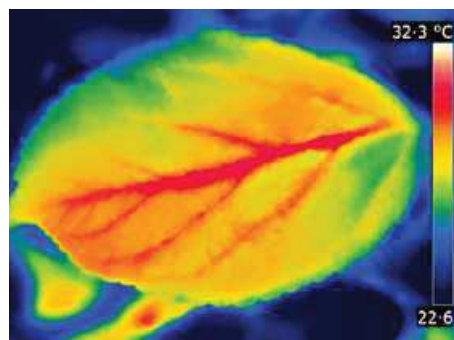
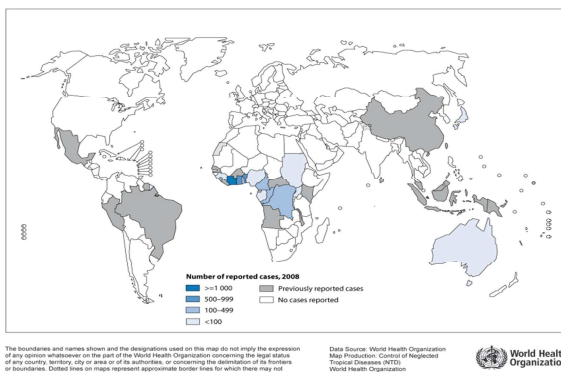


Figure 7. An infrared image of a leaf surface, taken in the field. The purpose of this example is to highlight the rapid change in condition, here temperature, over a very short distance. Averaging of climate can lead to loss of this important micro-variation, misleading our impression of suitable habitat. Reproduced from Caillon *et al.* (2014).

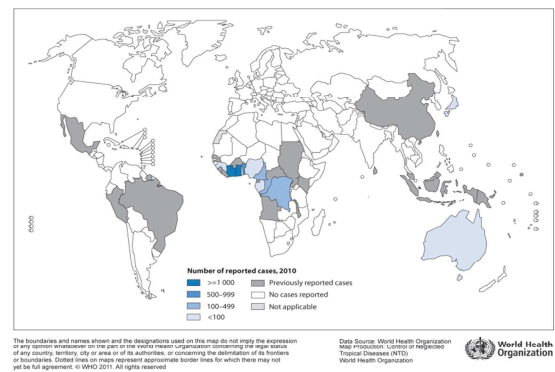
## Worldwide incidence, prevalence and pathology of Buruli ulcer

Buruli ulcer is an emergent neglected tropical disease, afflicting primarily poor agricultural communities in West and Central Africa. Cases have been persistently reported in Australia, Japan and French Guiana (an ultra-peripheral French territory in South America), and occasional cases have been reported in several other countries. As shown in Figure 8, the worldwide distribution falls mainly in West Africa, and globally between 5,000 and 10,000 people are infected per year (Johnson *et al.* 2005).

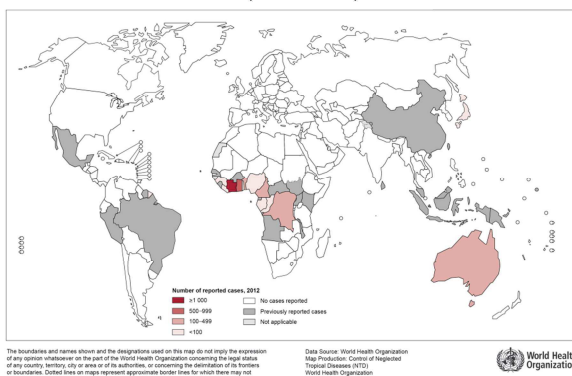
Distribution of Buruli ulcer, worldwide, 2008



Distribution of Buruli ulcer, worldwide, 2010



Distribution of Buruli ulcer, worldwide, 2012



Distribution of Buruli ulcer, worldwide, 2013

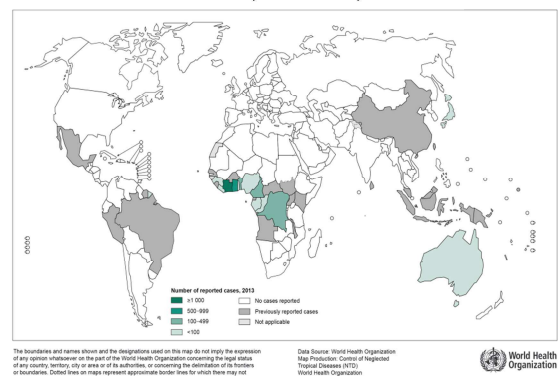


Figure 8. Global trends in Buruli ulcer reported cases worldwide in 2008, 2010, 2012 and 2014. The majority of cases occur in West and Central Africa, however cases have been consistently reported from French Guiana, Australia and Japan, with cases occasionally reported in many equatorial countries. Reproduced from the World Health Organisation.

Changes in the prevalence of Buruli ulcer in time are complex, and still not well understood. Incidence is known to respond to climate, particularly El Nino events, in French Guiana (Morris *et al.* 2014). In a long term temporal study in Akonolinga, Cameroon, the disease was found to expand according to local landscape characteristics; the distribution of agricultural land, wet land, and the local river, the Nyong, appear to drive the expansion of the disease agent in this region (Landier *et al.* 2014). Landscape and climate are not the only factors driving the prevalence of the Buruli ulcer, as complex social interactions with the changing landscape also play an important role (Giles-Vernick *et al.* 2014).

Preliminary infection by *M. ulcerans* results, after an unknown amount of time after the infection, in a small painless nodule. The normal evolution of the infection is to form a plague and oedema. Ulceration can eventually occur (Figure 9), and cause extensive permanent damage to the skin, muscle and bone. Once this has occurred the damage is normally permanent, leading to severe crippling deformities.

Treatment of *M. ulcerans* infection is normally through a combination of rifampin and streptomycin, which has proven effective (Etuaful *et al.* 2005). Though treatment is often provided free of charge, associated costs (travel, residence at the hospital, social stigma) are often prohibitive, particularly in poorer rural communities in Cameroon and Africa.





Figure 9. Infection by *M. ulcerans*. The disease has a complex progression, from nodule to ulcer. The incubation time is unknown, but estimated to be several months. Only the ulcerative stage is shown. Within the ulcer *M. ulcerans* normally resides in the undermined skin, which is red and inflamed, rather than the centre of the ulcer. Co-occurrence of other pathogens within the ulcer is remarkably rare. Reproduced from Johnson *et al.* (2005).

## **International response to Buruli ulcer emergence**

In response to the Buruli ulcer the French National Research agency (Agence nationale de la recherche, ANR) funded the project “Expansion and transmission of *Mycobacterium ulcerans* in changing environments: multidisciplinary studies” (EXTRA-MU) in 2012 (code ANR 11 CEPL 007 04 EXTRA-MU). This project brought together experts from a variety of disciplines, aiming to undertake an integrated research project studying the ecology and distribution of *M. ulcerans* and risk factors associated with human infection. This thesis has been financed primarily by the EXTRA-MU project, the Centre National de la Recherche Scientifique (CNRS), and Institut Pasteur. The thesis has been supported at the labs Maladies Infectieuses et Vecteurs : Écologie, Génétique, Évolution et Contrôle (MIVEGEC) and Jeunes équipes associées à l'IRD AtoMyc (JEAi AtoMyc) at the Institut de Recherche pour le Développement (IRD), and Territoires, Environnement, Télédétection et Information Spatiale (TETIS) at the Centre de coopération internationale en recherche agronomique pour le développement (CIRAD). The thesis is supported academically by the Ecole Doctorale Systèmes Intégrés en Biologie, Agronomie, Géosciences, Hydrosociences, Environnement (SIBAGHE) at the Université Montpellier 2. As part of ongoing collaboration with the Centre Pasteur du Cameroun, the EXTRA-MU project focused on Buruli ulcer in Cameroon.

Within this thesis much of the work is focused in Cameroon, however collaboration was also established, and the thesis partly supported by, the research group biodiversité et dynamiques des interactions dans des systèmes multi-hôtes-parasites. (BIOHOPSYs) at the Laboratoire d'Excellence: Centre d'étude de la biodiversité amazonienne (LabEx-CEBA). The LabEx CEBA focuses on the study and conservation of Amazonian biodiversity in French Guiana, a Buruli ulcer endemic region, and one of the goals of BIOHOPSYs is to evaluate the importance of biodiversity to human health. These three research projects, EXTRA-MU, JEAi AtoMyc and BIOHOPSYs, involve the work of several dozens of researchers and PhD students who have provided extensive support and advice throughout this research.

## **Evolution of *M. ulcerans***

The pathogen that is the focus of this large research effort, *M. ulcerans*, has proven to be a particularly difficult microbe to study. This difficulty arises due to its unique ecology and evolutionary history.

*M. ulcerans* appears to have evolved from *Mycobacterium marinum* (Stinear *et al.* 2000). *M. marinum* is a fast growing aquatic pathogen, commonly infecting fish and humans who come in close

contact with aquatic environments. Human infection appears to result from cuts and abrasions of the skin, resulting in intracellular infection of host cells. *M. marinum* infection results in nodules (Figure 10, Lewis *et al.* 2003) that are normally easily treatable with standard antimicrobial drugs. This pathogen also appears to be able to survive in the environment outside the host, producing pigments to protect itself from UV light.



Figure 10. Result of human infection by *M. marinum*. Reproduced from Lewis *et al.* (2003). ©2003 by the Infectious Diseases Society of America.

*M. marnium* appears to have undergone a population bottleneck, acquired a plasmid (pMUM001) and experienced extensive gene inactivation due to the acquisition of several copies of insertion sequences, which may have been acquired from *Streptomyces*, an occasionally pathogenic genus of soil bacteria (Stinear *et al.* 2007). This reduction in genome size and increased sensitivity to environmental factors is characteristic of bacteria that are adapting to a new environment, suggesting that *M. ulcerans* has recently adapted to occupy a new ecological niche, different to *M.*

*marinum* (Parkhill *et al.* 2003) The result is a bacterium, *M. ulcerans*, with an extremely slow growth rate (72 hour doubling time), high sensitivity to UV light and sensitivity to contamination in culture. One of the most notable features of *M. ulcerans* is the production of a virulence factor, mycolactone. Mycolactone is a large lipid molecule, structurally similar to the macrocyclic triene rapamycin from the soil pathogen *Streptomyces*, which is also involved in immunosuppression (Coutanceau *et al.* 2007). Mycolactone induces apoptosis in eukaryotic cells (George *et al.* 2000), which leads to local immunosuppression through destruction of phagocytes. The painless nature of the Buruli ulcer nodules is a result of functional suppression of the nerve cells by mycolactone (Coutanceau *et al.* 2007).

### **Environmental distribution of *M. ulcerans***

The ecological niche and environmental distribution of *M. ulcerans* remains unknown, but previous work has associated the bacteria with certain types of environment. Buruli ulcer is known to be associated with disturbed aquatic environments in Africa (Merritt *et al.* 2010), though not Australia, where its associations are less clear but appear to be urban (Johnson *et al.* 2007). *M. ulcerans* is expected to show a similar association aquatic environments, and the majority of previous field work has focused on aquatic sampling, however this has recently been questioned (below). Wagner *et al.* (2008) associated Buruli ulcer (*M. ulcerans* present in humans) to non-urban landscapes, with a high wetness index, and Williamson *et al.* (2008) found that the bacterium has a distribution wider than the disease (i.e., *M. ulcerans* can be present in the environment without infecting humans). Previous work has found that landscape, particularly aquatic sites that are lentic or lotic, influence *M. ulcerans* distribution. *M. ulcerans* is more commonly detected in swamps (lentic systems) than rivers (lotic systems) in a study in Ghana (Figure 11, Benbow *et al.* 2013, McIntosh *et al.* 2014, García-Peña *et al.* in preparation [Chapter 6]).

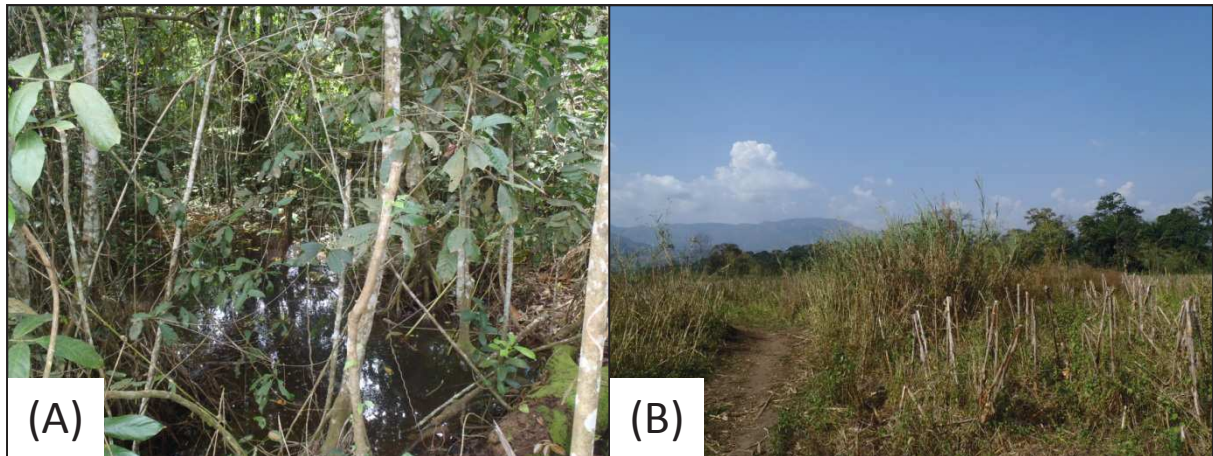


Figure 11. Examples of landscapes typically associated with *M. ulcerans* presence (A). This slow flowing, stagnant water is in a lowland swamp in Bankim, Cameroon. (B) represents a landscape typically expected to have low levels of *M. ulcerans* presence. Here, mountainous terrain creates small rivers and streams rather than swamps, also in Bankim. © Kevin Carolan 2012.

Due to the difficulty in culturing *M. ulcerans*, environmental detection often uses qPCR to detect the presence of the bacterium. As discussed above, this raises issues of detection bias. *M. ulcerans* DNA has been detected in a surprising variety of environmental samples, including aquatic insects, biofilms, crustaceans, detritus, fish, frogs, mammals, marsupials, soil, snails, water and worms (Portaels *et al.* 1999, Stinear *et al.* 2000, Portaels *et al.* 2001, Eddyani *et al.* 2004, Marsollier *et al.* 2004, Trott *et al.* 2004, Kotlowski *et al.* 2004, Johnson *et al.* 2007, Fyfe *et al.* 2007, Williamson *et al.* 2008, Portaels *et al.* 2008, Fyfe *et al.* 2010, Marion *et al.* 2010, Roche *et al.* 2013, Carson *et al.* 2014, Morris *et al.* 2014, Garchitorena *et al.* 2014). This combination of a widely distributed bacterium which appears to be present in a wide variety of species is surprising, given the apparently reduced ecological niche implied by the genome.

A discussion on the ecology of *M. ulcerans* deserves to note the unusual case of Australia. The ecological distribution of *M. ulcerans* in Australia is notably different from other regions. In Australia the bacterium is detected in possums. These large arboreal marsupials are common pests in urban habitats, and infection in possums is closely associated with human cases of the disease (Carson *et al.* 2014). Outbreaks of the Buruli ulcer appear to be smaller, with temporary transient outbreaks over a short distance (approximately 1 or 2 kilometres), compared to Africa, where the disease tends to be chronic to a particular area, infecting people across tens of kilometres (Johnson *et al.* 2005). The source of infection for possums is unclear, however they have been proposed as key hosts and potential sentinels of the disease. The reasons for this distinctly different ecology are unclear. Two hypotheses have been proposed. *M. ulcerans* appears to have been introduced to Gippsland (Eastern Australia) and moved westward towards Melbourne; potentially the bacterium has further

adapted and evolved, or it is now occupying a previously unoccupied section of its ecological niche (it had not encountered possums in Africa, but may be able to infect them). This different ecological context would then lead to different distributions of suitable habitat for the bacterium, as in Figure 12. The cause is unclear, and will be the subject of future work.

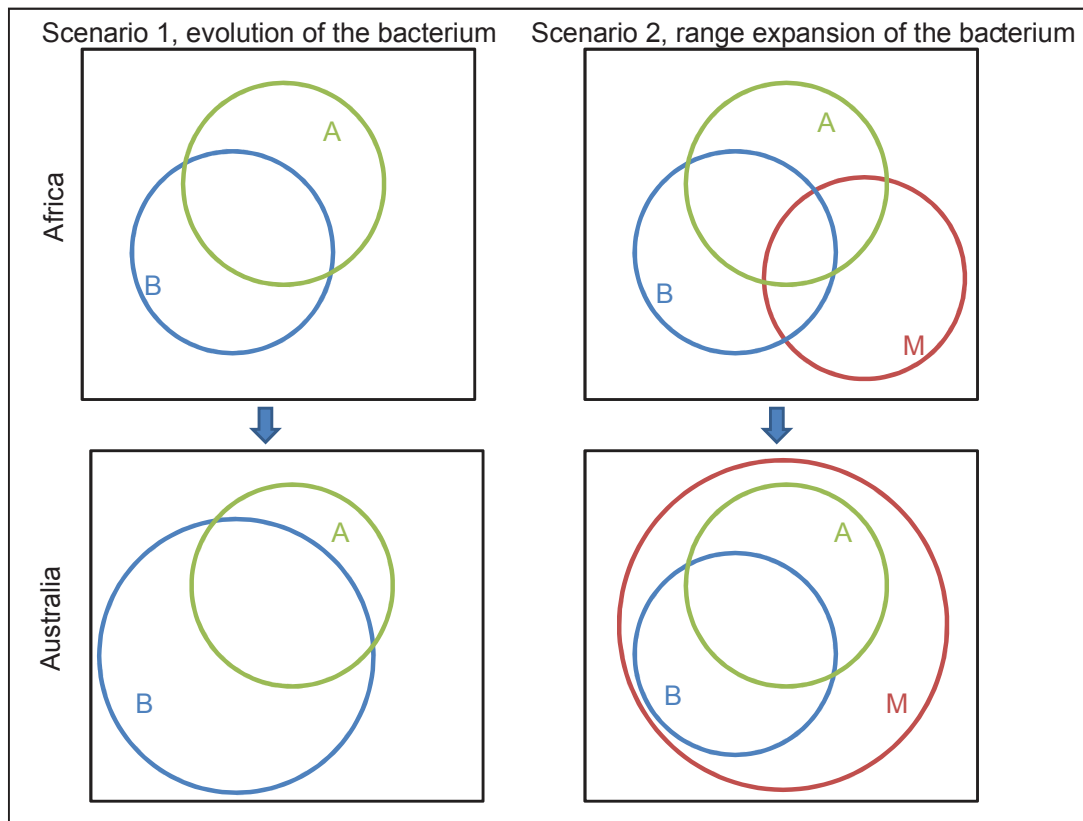


Figure 12. Potential scenarios explaining the different distribution of the disease in Australia. Here, the range of suitable combinations of Biotic and Abiotic variables are represented by  $B \cap A$ . In Scenario 1, the bacterium has adapted to a new biotic reservoir on being introduced to Australia from Africa, changing the size of  $B \cap A$ , and resulting in a different ecological distribution. The second scenario represents the niche with three determinants, where  $M$  represents movement limitations. Here, the bacteria does not undergo evolution or adaption, but the expansion of  $M$  increases the size of  $B \cap A \cap M$ , resulting in a different species distribution. The exact cause of the difference remains to be identified, and will have implications for the applicability of studies from one region to another.

### Routes by which *M. ulcerans* infects humans

The route of transmission is unknown. The bacterium is known to be environmentally acquired, and without human to human transmission, and several modes of environmental transmission have been proposed. Aquatic insects have been proposed as likely vectors of the Buruli ulcer (Marsollier *et al.* 2002), however the likelihood of this route is debated (Benbow *et al.* 2008). Aquatic insects of the Families Naucoridae and Belostomadiae (Figure 13), known to very occasionally accidentally bite humans, are able to transmit the bacteria from snails to mice in the lab, resulting in Buruli ulcer like

symptoms in the mice (Marsollier *et al.* 2002), and are common in aquatic environments associated with the disease. They are considered as unlikely vectors, due to the accidental nature of their biting humans, and the lack of any observed spatial or temporal correlation between their abundance and the prevalence of Buruli ulcer. The potential role of these insects is explored further in the fourth chapter of this thesis.

An alternative hypothesis is that of infection through direct contamination, as is the case for most *M. marinum* infections. This supposes that damage to the skin, through cuts and abrasions, opens an opportunity for the bacterium to infect the wound via physically being deposited there during aquatic exposure. However, recent lab studies of direct inoculation of *M. ulcerans* onto wounded flesh in Guinea pigs failed to result in infection (Williamson *et al.* 2014). This would suggest that direct contamination is unlikely.

In Australia aerosolization of the bacterium has been proposed as a potential method in infection, or dispersal (Hayman 1991, Ross *et al.* 1997). This was originally proposed following observations of a uniquely coastal distribution of the disease in Australia, however the sensitivity of the bacterium to UV light suggested this route was unlikely. Recently the potential importance of this route has been revived, following the discovery of the importance of free living amoebae as hosts of *M. ulcerans*, suggesting that the amoebae may provide the protection needed to survive in the aerosolized environment (Amissah *et al.* 2014).

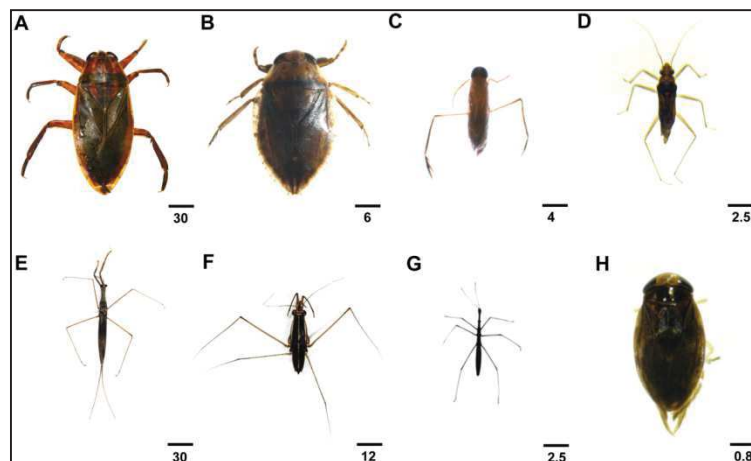


Figure 13. Examples of aquatic insects (Order Hemiptera) collected from endemic regions of Cameroon, of the Family Belostomatidae, genus' *Appasus* (A) and *Lethocerus* (B), the Families Notonectinae (C) and Mesoveliidae (D), the Family Nepidae, genus *Ranatrinae* (E), the Family Gerridae (F), the Family Hydrometridae (G), and the Family Corixidae genus *Micronectinae* (H). Scale bars in mm. Reproduced from Marion *et al.* 2010.

## **Overview of the thesis**

In summary, *M. ulcerans* is an emerging neglected tropical pathogen afflicting between 5,000 to 10,000 people per year. The etiology of the disease, including the pathogen's environmental reservoir and mode of transmission, are unknown. The goal of this thesis is to describe the ecological niche at the landscape scale to assist in the production of hazard maps, to explore the aquatic insect model of transmission in order to identify the likelihood of this route of transmission. These goals are aimed to assist in the control of the Buruli ulcer.

In order to use the variety of data sources effectively, the Chapter 2 address questions related to model performance and sample size. Chapter 3 then applies these to the ecological niche of *M. ulcerans*, while Chapter 4 applies them to the proposed vectors, and the final chapter, Chapter 5, synthesises the results and concludes the thesis.



## References

- Amissah, N., *et al.* "Investigating the Role of Free-living Amoebae as a Reservoir for *Mycobacterium ulcerans*" *PLoS Neglected Tropical Diseases* (2014) DOI: 10.1371/journal.pntd.0003148
- Austin, M., "Species distribution models and ecological theory: a critical assessment and some possible new approaches." *Ecological modelling* 200.1 (2007): 1-19.
- Benbow, E., *et al.* "Aquatic Macroinvertebrate Assemblages of Ghana, West Africa: Understanding the Ecology of a Neglected Tropical Disease." *Ecohealth* (2013): 1-16.
- Benbow, E., *et al.* "Aquatic invertebrates as unlikely vectors of Buruli ulcer disease." *Emerging infectious diseases* 14.8 (2008): 1247.
- Bhatt, S., *et al.* "The global distribution and burden of dengue." *Nature* (2013): 504–507 doi:10.1038/nature12060
- Carson, C., *et al.* "Potential Wildlife Sentinels for Monitoring the Endemic Spread of Human Buruli Ulcer in South-East Australia." *PLoS neglected tropical diseases* 8.1 (2014): e2668.
- Caillon, R., *et al.* "Warming decreases thermal heterogeneity of leaf surfaces: implications for behavioural thermoregulation by arthropods." *Functional Ecology* (2014).
- Coutanceau, E., *et al.* "Modulation of the host immune response by a transient intracellular stage of *Mycobacterium ulcerans*: the contribution of endogenous mycolactone toxin." *Cellular microbiology* 7.8 (2005): 1187-1196.
- Coutanceau, E., *et al.* "Selective suppression of dendritic cell functions by *Mycobacterium ulcerans* toxin mycolactone." *The Journal of experimental medicine* 204.6 (2007): 1395-1403.
- Eddyani, M., *et al.* "Potential role for fish in transmission of *Mycobacterium ulcerans* disease (Buruli ulcer): an environmental study." *Applied and environmental microbiology* 70.9 (2004): 5679-5681.
- Elith J., *et al.* "Novel methods improve prediction of species' distributions from occurrence data." *Ecography* 29.2 (2006): 129-151.
- Elith, J., *et al.* "The art of modelling range-shifting species." *Methods in ecology and evolution* 1.4 (2010): 330-342.
- Etuaful, S., *et al.* "Efficacy of the combination rifampin-streptomycin in preventing growth of *Mycobacterium ulcerans* in early lesions of Buruli ulcer in humans." *Antimicrobial agents and chemotherapy* 49.8 (2005): 3182-3186.
- Fyfe, J., *et al.* "Development and application of two multiplex real-time PCR assays for the detection of *Mycobacterium ulcerans* in clinical and environmental samples." *Applied and environmental microbiology* 73.15 (2007): 4733-4740.
- Fyfe, J., *et al.* "A major role for mammals in the ecology of *Mycobacterium ulcerans*." *PLoS neglected tropical diseases* 4.8 (2010): e791.

Garchitorena, A., *et al.* "Mycobacterium ulcerans Ecological Dynamics and Its Association with Freshwater Ecosystems and Aquatic Communities: Results from a 12-Month Environmental Survey in Cameroon." *PLoS neglected tropical diseases* 8.5 (2014): e2879.

George, K., *et al.* "A Mycobacterium ulcerans toxin, mycolactone, causes apoptosis in guinea pig ulcers and tissue culture cells." *Infection and immunity* 68.2 (2000): 877-883.

Gething, P., *et al.* "Climate change and the global malaria recession." *Nature* 465.7296 (2010): 342-345.

Giles-Vernick, T., *et al.* "The puzzle of Buruli ulcer transmission, ethno-ecological history and the end of "love" in the Akonolinga district, Cameroon." *Social Science & Medicine* (2014). DOI: 10.1016/j.socscimed.2014.03.008

Glass, G., *et al.* "Environmental risk factors for Lyme disease identified with geographic information systems." *American Journal of Public Health* 85.7 (1995): 944-948.

Gravel, D., *et al.* "Reconciling niche and neutrality: the continuum hypothesis." *Ecology Letters* 9.4 (2006): 399-409.

Guisan, A. *et al.* "Generalized linear and generalized additive models in studies of species distributions: setting the scene." *Ecological modelling* 157.2 (2002): 89-100.

Hay, S., *et al.* "Big data opportunities for global infectious disease surveillance." *PLoS medicine* 10.4 (2013): e1001413.

Haydon, D., *et al.*, "Integrating epidemiology into population viability analysis: managing the risk posed by rabies and canine distemper to the Ethiopian wolf." *Conservation Biology* 16.5 (2002): 1372-1385.

Hayman, J., "Postulated epidemiology of Mycobacterium ulcerans infection." *International journal of epidemiology* 20.4 (1991): 1093-1098.

Hirzel, A., *et al.* "Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data?." *Ecology* 83.7 (2002): 2027-2036.

Hutchinson, G.E. (1957). "Concluding remarks" (PDF). *Cold Spring Harbor Symposia on Quantitative Biology* 22 (2): 415–427.

Johnson, P. D., *et al.* "Mycobacterium ulcerans in mosquitoes captured during outbreak of Buruli ulcer, southeastern Australia." *Emerg Infect Dis* 13.11 (2007): 1653-1660.

Kearney, *et al.* "Mapping the fundamental niche: physiology, climate, and the distribution of a nocturnal lizard." *Ecology* 85.11 (2004): 3119-3131.

Kotlowski, R., *et al.* "One-tube cell lysis and DNA extraction procedure for PCR-based detection of Mycobacterium ulcerans in aquatic insects, molluscs and fish." *Journal of medical microbiology* 53.9 (2004): 927-933.

Landier, J., *et al.* "Spatiotemporal variations and landscape-associated risk of Buruli ulcer in Akonolinga, Cameroon" *PLoS neglected tropical diseases* (2014) DOI: 10.1371/journal.pntd.0003123

Lewis, F., *et al.*, "Fish tank exposure and cutaneous infections due to *Mycobacterium marinum*: tuberculin skin testing, treatment, and prevention." *Clinical infectious diseases* 37.3 (2003): 390-397.

Lockwood, J., *et al.* "The role of propagule pressure in explaining species invasions." *Trends in Ecology & Evolution* 20.5 (2005): 223-228.

Manel, S., *et al.* "Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird." *Ecological Modelling* 120.2 (1999): 337-347.

Marion, E., *et al.* "Seasonal and regional dynamics of *M. ulcerans* transmission in environmental context: deciphering the role of water bugs as hosts and vectors." *PLoS neglected tropical diseases* 4.7 (2010): e731.

Marsollier, L., *et al.* "Aquatic insects as a vector for *Mycobacterium ulcerans*." *Applied and environmental microbiology* 68.9 (2002): 4623-4628.

Marsollier, L., *et al.* "Aquatic plants stimulate the growth of and biofilm formation by *Mycobacterium ulcerans* in axenic culture and harbor these bacteria in the environment." *Applied and environmental microbiology* 70.2 (2004): 1097-1103.

McIntosh, M., *et al.* "Associations between *Mycobacterium ulcerans* and aquatic plant communities of West Africa: implications for Buruli ulcer disease." *EcoHealth* (2014): 1-13.

Meentemeyer, R., *et al.* "Early detection of emerging forest disease using dispersal estimation and ecological niche modeling." *Ecological Applications* 18.2 (2008): 377-390.

Merritt, R., *et al.* "Ecology and transmission of Buruli ulcer disease: a systematic review." *PLoS neglected tropical diseases* 4.12 (2010): e911.

Morris, A., *et al.* "First Detection of *Mycobacterium ulcerans* DNA in Environmental Samples from South America." *PLoS neglected tropical diseases* 8.1 (2014): e2660.

Morris, Aaron, *et al.* "Complex temporal climate signals drive the emergence of human water-borne disease." *Emerging Microbes & Infections* 3.8 (2014): e56.

Palomino, J. C., *et al.* "Effect of oxygen on growth of *Mycobacterium ulcerans* in the BACTEC system." *Journal of clinical microbiology* 36.11 (1998): 3420-3422.

Parkhill, J., *et al.* "Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*." *Nature genetics* 35.1 (2003): 32-40.

Peterson, A. *et al.* "Predicting distributions of Mexican birds using ecological niche modelling methods." *Ibis* 144.1 (2002): E27-E32.

Peterson, A. *et al.* "Using Ecological-Niche Modeling to Predict Barred Owl Invasions with Implications for Spotted Owl Conservation." *Conservation Biology* 17.4 (2003): 1161-1165.

Portaels, F., *et al.* "Insects in the transmission of *Mycobacterium ulcerans* infection." *The Lancet* 353.9157 (1999): 986.

Portaels, F., *et al.* "Mycobacterium ulcerans in wild animals." *Revue scientifique et technique (International Office of Epizootics)* 20.1 (2001): 252.

Portaels, F., *et al.* "First cultivation and characterization of *Mycobacterium ulcerans* from the environment." *PLoS neglected tropical diseases* 2.3 (2008): e178.

Roche, B., *et al.* "Identifying the Achilles heel of multi-host pathogens: the concept of keystone 'host' species illustrated by *Mycobacterium ulcerans* transmission." *Environmental Research Letters* 8.4 (2013): 045009.

Ross, B. C., *et al.* "Detection of *Mycobacterium ulcerans* in environmental samples during an outbreak of ulcerative disease." *Applied and environmental microbiology* 63.10 (1997): 4135-4138.

Segurado, P., and Araujo M., "An evaluation of methods for modelling species distributions." *Journal of Biogeography* 31.10 (2004): 1555-1568.

Soberón, J., and Nakamura M., "Niches and distributional areas: concepts, methods, and assumptions." *Proceedings of the National Academy of Sciences* 106.Supplement 2 (2009): 19644-19650.

Simard, F., *et al.* "Ecological niche partitioning between *Anopheles gambiae* molecular forms in Cameroon: the ecological side of speciation." *BMC ecology* 9.1 (2009): 17.

Stinear, T., *et al.* "Identification of *Mycobacterium ulcerans* in the environment from regions in Southeast Australia in which it is endemic with sequence capture-PCR." *Applied and environmental microbiology* 66.8 (2000): 3206-3213.

Stinear, T., *et al.* "Reductive evolution and niche adaptation inferred from the genome of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer." *Genome research* 17.2 (2007): 192-200.

Tilman, D. and Kareiva P, *Spatial ecology: the role of space in population dynamics and interspecific interactions*. Vol. 30. Princeton University Press, 1997.

Trott, K., *et al.* "Characterization of a *Mycobacterium ulcerans*-like infection in a colony of African tropical clawed frogs (*Xenopus tropicalis*)." *Comparative medicine* 54.3 (2004): 309-317.

Tran, A., *et al.* "Using remote sensing to map larval and adult populations of *Anopheles hyrcanus* (Diptera: Culicidae) a potential malaria vector in Southern France." *International Journal of Health Geographics* 7.1 (2008): 9.

Wagner, T., *et al.* "A landscape-based model for predicting *Mycobacterium ulcerans* infection (Buruli ulcer disease) presence in Benin, West Africa." *EcoHealth* 5.1 (2008): 69-79.

Williamson, H., *et al.* "Distribution of *Mycobacterium ulcerans* in Buruli ulcer endemic and non-endemic aquatic sites in Ghana." *PLoS neglected tropical diseases* 2.3 (2008): e205.

Williamson, H., *et al.* "Mycobacterium ulcerans Fails to Infect through Skin Abrasions in a Guinea Pig Infection Model: Implications for Transmission." *PLoS neglected tropical diseases* 8.4 (2014): e2770.

Zuur, A., *et al.* *Mixed effects models and extensions in ecology with R*. Springer, 2009.

## **Chapter 2: How many samples do we need to describe the spatial distribution of an emerging pathogen?**

## **Foreword.**

The following chapter is currently under review as a paper to the Journal of Epidemiology and Infection, titled “How many samples do we need to describe the spatial distribution of an emerging pathogen?”, by Kevin Carolan, Danny Lo Seen and Jean-François Guégan.

This article originated from a need to appropriately use information from multiple databases. Within the EXTRA-MU and BIOHOPSYS projects I collaborated with three researchers conducting work in separate environmental contexts. The goal of this project was to evaluate how the performance of each model would be influenced by the different sample sizes used by each of the researchers.

Within the EXTRA-MU project Andres Garchitorrena conducted 12 months of sampling in Cameroon (Figure 14), visiting two Buruli ulcer endemic sites, Akonolinga and Bankim, and Estelle Marion conducted sampling in Pobè, Benin. As part of the LabEx-CEBA BIOHOPSYS programme, and supported by Bournemouth University, Arron Morris undertook a year of field work in French Guiana. These four sites, Akonolinga, Bankim, Benin, and French Guiana received similar (though not identical) sampling protocols and had dramatically different sample sizes. Though the following article is written in terms of the sample size needed for pathogens, the principles are generally applicable to most species.

Grateful thanks are extended to the staff of IRD; Andres Garchitorrena, Benjamin Roche, and Annelise Tran of CIRAD, for critical review of the manuscript and insightful discussions on the subject.

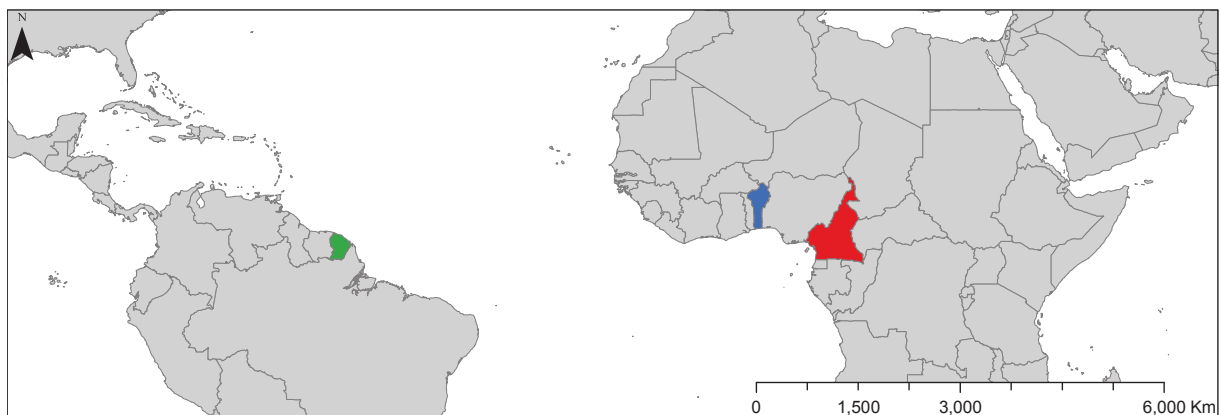


Figure 14. Three endemic regions studied herein. Field work by Aaron Morris was conducted in French Guiana (Green), work by Estelle Marion was in Benin (Blue) and Andres Garchitorrena worked in Cameroon (Red).

## **Abstract**

Answering questions of where the pathogens are is often the first step in the control of diseases. Ecological niche models are widely used to describe the spatial distribution of macro-organisms, such as birds or mammals, and in the last decade their use has increased considerably for the study of disease and emerging environmental pathogens. Here we explore the sample size needed for good performance, and discuss challenges unique to modelling the distribution of environmental zoonoses and environmentally-acquired microbes. We create several virtual species and model their ecological niches and spatial distribution with the software Maxent. We find that the sample size needed for a niche model cannot be determined without consideration of the ecological context in which a given pathogenic species is distributed. These results highlight the need to appropriately limit the extent of the study, with political boundaries and health districts, commonly used in ecological epidemiology, rarely being appropriate. Selection of the extent of the study should be based on an ecological assumption or hypothesis about the pathogen as a species in ecology, which can provide a guideline for future studies.

## **Introduction**

To aid control of diseases it is often important to understand the spatial distribution of pathogenic zoonotic species, where incorrect expectations can have a high cost to society. How can we shape our predictions of the spatial distribution of these environmental pathogens to be as close to reality as possible? A potential solution lies in the field of ecological niche modelling (ENM), which describes the range of environmental conditions within which a species can maintain a population (Hutchinson 1957, Soberón and Nakamura 2009). ENM is often used to inform conservation efforts (Peterson and Robins 2003), to describe species potential range (Meentemeyer *et al.* 2004), and is increasingly being used to describe species of public health concern, having been applied to pathogens or their vectors (Ayala *et al.* 2009, Mullins *et al.* 2011, Daszak *et al.* 2013, Hay *et al.* 2013). Given the utility of ENM, and its potential benefit to public health, how do we go about constructing these models? How do we know over what area to conduct the study, how many samples are needed to construct an appropriate niche model, and how do we evaluate the performance of the models?

In this paper we discuss the complex relationships between sample size, species ecology, and region of study, and demonstrate how these factors influence the performance of the model, all of which contribute to an understanding of the spatial distribution of pathogens.

To begin with, a common source of confusion is that the concept of sample size differs between epidemiological and ecological literature. As a hypothetical example; suppose we collect 500



individual members of an insect species suspected to be vectors of a pathogenic agent, and we collect these samples in 5 ecologically unique regions. Epidemiologically we have 500 samples to test for the presence of the pathogen. Ecologically, this is a sample size of 5, and the 500 individuals change from being samples, to being abundance data. This simple change in mindset between epidemiology and ecology has important implications for ENM of pathogens, as discussed herein.

In order to conduct ENM we first select the environmental variables that we suspect describe the ecological niche (Hutchinson 1957). The niche is often represented in a simplified manner with a type of Venn diagram termed the Biotic-Abiotic-Migratory (BAM) diagram (Soberón and Nakamura 2009), as shown in Figure 15. We use the structure of the BAM diagram to represent our hypothesis about the importance each of the different environmental variables, which influences our ability to accurately model the niche; not all structures are equally easy to describe (Saupe *et al* 2012). Examples of such hypotheses are found in Figures 15 and 17.

The region in Figure 15 noted  $B \cap A$  is the set of suitable habitat for the species; the purpose of ENM is to describe this.  $B \cap A \cap M^c$  represents the area of suitable habitat ( $B \cap A$ ) that is currently inaccessible ( $M^c$ ) and may be vulnerable to future invasion by the species. If the species is a vector or pathogen,  $B \cap A \cap M^c$  represents a region of future concern for public health authorities. The structure of BAM diagram can represent many different scenarios; in Figure 16A biotic variables dominate  $B \cap A$ , so that  $B \cap A = B$  and the species' suitable habitat is limited entirely by biotic variables. An example of such a species is the pathogen *Phytophthora ramorum*, the causative agent of Sudden Oak Death, which is limited to a greater degree by the distribution of its host species rather than the distribution of abiotic variables such as temperature (Meentemeyer *et al* 2004).

In a second example, Figure 16B, the species is limited by its migration ability. Examples of this situation can include the malaria plasmodium, *Plasmodium vivax*. This pathogen was eradicated from southern France in 1943 following the drying of marshes, which destroyed vector habitat, and extensive use of quinine, which is toxic for *P. Vivax*. However, the environment is now suitable for the vector, and it currently poses a threat of re-invasion along with the plasmodium, potentially leading to a new outbreak of malaria in France (Tran *et al* 2008). Within the BAM, this is the region  $B \cap A^c \cap M$ ; where high quinine use made the region abiotically unsuitable, but following depletion of the chemical the habitat suitability increased, and returned to  $B \cap A \cap M$ .

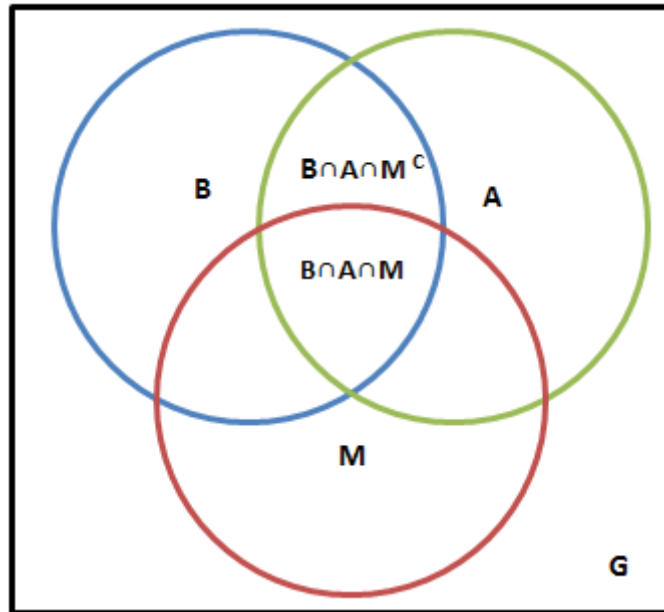


Figure 15. The Biotic-Abiotic-Movement (BAM) diagram, a 2D representation of the ecological niche, useful for understanding the ecology and distribution of emerging pathogens. The region **G** indicates all combinations of environmental variable values within the area of interest for a particular pathogen. A subset of this contains the biotic variables that are suitable for the pathogen, **B**, indicated in the blue circle. This can include, for example, distribution of suitable hosts, or the vulnerability of hosts to infection. The subset of **G** noted by **M** (migration) is the region of environmental variables the species can actually access over the time period of interest, such as a connected population of susceptible humans, or a region accessible by a malaria vector that is not isolated by climate. The region noted by **A** contains suitable abiotic values, which are important for certain environmental pathogens such as anthrax or tetanus, which only survive in certain soil types. The region **B ∩ A** (read as “intersection of **B** and **A**”) is the suitable biotic and abiotic conditions within which a pathogen or vector can maintain a population of individuals. This region is divided into **B ∩ A ∩ M**, which is accessible and suitable, and is where we expect to find the species (though meta-population dynamics and species detectability issues will influence our ability to detect it). The region of **B ∩ A ∩ M** that coincides with a human population is where disease occurs, one might imagine a third circle, **H**, that represents this region. The notation **M<sup>c</sup>** is read as “the compliment of **M**”, the region of habitat suitable for the pathogen not overlapped by **M**, and is currently inaccessible to the species. This region, **B ∩ A ∩ M<sup>c</sup>** is often of great concern for public health authorities, as it represents areas of potential future epidemics.

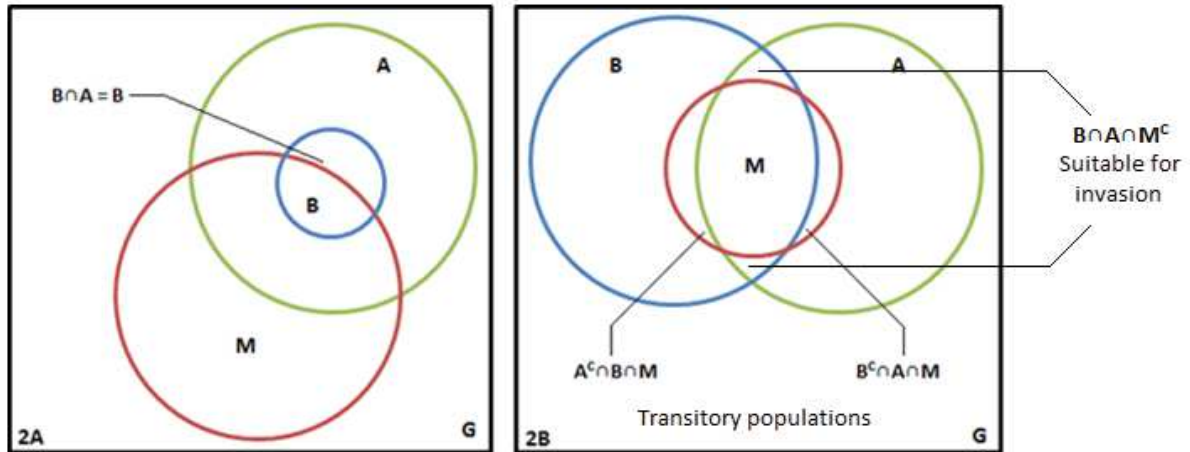


Figure 16. Different structures of the BAM diagram. The structure of the BAM diagram changes for different species, or for the same species in a different region of interest. For example, the distribution of malaria vectors will be determined by abiotic environmental conditions in one region (such as rainfall) and by biotic conditions (such as competition) in other regions. In 2A biotic factors are a greater limiting factor on the size of  $B \cap A \cap M$  than abiotic factors, such that  $B \cap A = B$ . This is the type of system traditionally studied in epidemiology, where a pathogen is confined to the body of a host (intra-host ecology). In 2B this species (be it pathogen or vector) also has two regions accessible to, but unsuitable for, the species, marked  $B^c \cap A \cap M$  and  $B \cap A^c \cap M$  (where  $B^c$  and  $A^c$  are the complement of biotic and complement of abiotic respectively, the areas of unsuitable environmental conditions) where transitory populations can be found. These transitory populations have important implications for disease control. For example, a given bacterium may switch between commensal or pathogenic life styles under situations of environmental stress (such as *Bacillus anthracis*, the causative agent of anthrax). A second example is that of insects, acting as disease vectors, straying into these habitats. The choice of best method to control the populations of such vectors will change depending on the structure of the BAM diagram.

In Figure 16, we see that the region of suitable habitat,  $B \cap A$ , is larger in 16B than 16A, assuming  $G$  in both diagrams is equivalent. This results in a larger amount of information to be described by the ENM, which effects model performance, as discussed below. The specialist species, with the smaller ecological hypervolume, will require fewer samples to describe than a more generalist species with a larger hypervolume (Hernandez *et al* 2006). However, this effect is slightly more subtle than it first appears; as Figure 17 illustrates. As our region of interest changes from  $G_1$  (a small scale local study of the vector or pathogen) to  $G_2$  (a large or global scale) the volume of  $B \cap A$  also increases. This causes the species to appear to be a specialist for one region or variable, and generalist for another, and this increase in the size of  $B \cap A$  should also increase the amount of information we need to explain.

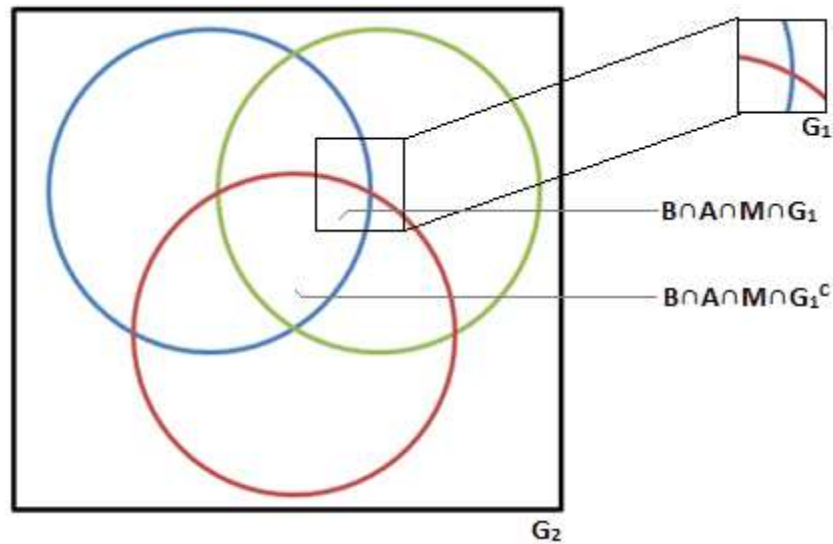


Figure 17. Subsetting BAM diagram representing the effect of changing the region of interest. Region  $G_1$  is a subset of  $G_2$ . For a study bound by the limits of  $G_1$  abiotic factors are not a limit on species distribution, and a model constructed in this region will not account for their effect in  $B \cap A \cap M \cap G_1^c$ . Though it is simpler to interpret increasing  $G$  as an increase in the spatial area of the study,  $G$  also increases if we increase the number of environmental variables studied.

The purpose of these examples is to demonstrate that the relative importance of different environmental variables in determining the distribution of an environmental pathogen will change as we change species or the geographic region (or ecological volume) we are studying. Change in these ecological processes drives changes in the distribution of pathogens, causing the emergence of many modern zoonotic and environmentally acquired diseases. This change influences our ability to construct an accurate ENM; as the size of the region  $B \cap A$  of the BAM diagram changes the amount of information needed to describe that volume will also change, and we can facilitate understanding of this change with the BAM diagram (VanDerWal et al. 2009, Barve et al. 2011, Elith et al. 2011).

Herein, we explore the effect sample size ( $u$ ), niche volume ( $V$ ) and region of interest ( $G$ ) have on our ability to model the ecological niche of such a zoonotic or environmentally-acquired species. We evaluate model performance with the area under the curve of the receiver operating characteristic (AUC), a metric originally inherited from the medical literature (Matthews *et al* 1990). We theorised that by placing models in an ecological context, using the BAM diagram, we could explain how the required sample size changes in different environmental conditions.

We focus on a study of modelling methods that are limited to presence-only data. This is a common limitation in ecological epidemiology, as absence data is often difficult to gather, or underestimated in importance. While absence data can be extremely informative ecologically, it is often considered as methodological fault in epidemiology. To construct these presence-only models

we use maximum entropy modelling (Phillips, Dudík and Schapire 2004, Dudík, Phillips and Schapire 2007), in the software Maxent (Phillips, Anderson and Schapire 2006) as this is one of the most commonly used presence-only modelling methods in the literature (Anderson *et al* 2006). We discuss the utility of AUC as an index of Maxent model performance; exploring the effect  $u$ ,  $V$  and  $G$  have on model AUC, and how the ability of model AUC as an indicator of true performance changes.

## Materials and Methods

To explore the effect sample size ( $u$ ), niche volume ( $V$ ) and region of interest ( $G$ ) have on model performance and model AUC we divide this study into three questions. We also discuss the utility of AUC as an index of model performance and how the relationship between AUC and true performance changes. The three questions asked are as follows:

- 1) How does model performance change as sample size changes?

Ecologically unique sample sites range in number from 5 to 50 at intervals of 5; this range was selected to correspond to most previous environmental studies. Sample size is increased, and result on the model AUC is explored.

- 2) How does model performance change as the size of the species ecological niche (niche breadth) changes?

The volume of the species ecological niche ( $B \cap A$ ) is the volume of environmental variables within which it can be found. This is increased as described below; 5 different virtual species are used with increasing niche breadth.

- 3) How does model performance change as the region of interest changes?

The size of the set of environmental variable combinations under study ( $G$ ) is increased and the resulting effect on AUC and model performance is recorded. This is achieved simply by limiting our study area to different spatial extents, noted  $G_1$ , to  $G_5$ , shown in Figure 18. It should be recalled that this increase could also be achieved by increasing the number of environmental variables studied. An immediate criticism of using these 5 bounding regions is that they have no biological justification. We invite with this criticism, and the importance is discussed in the conclusion.

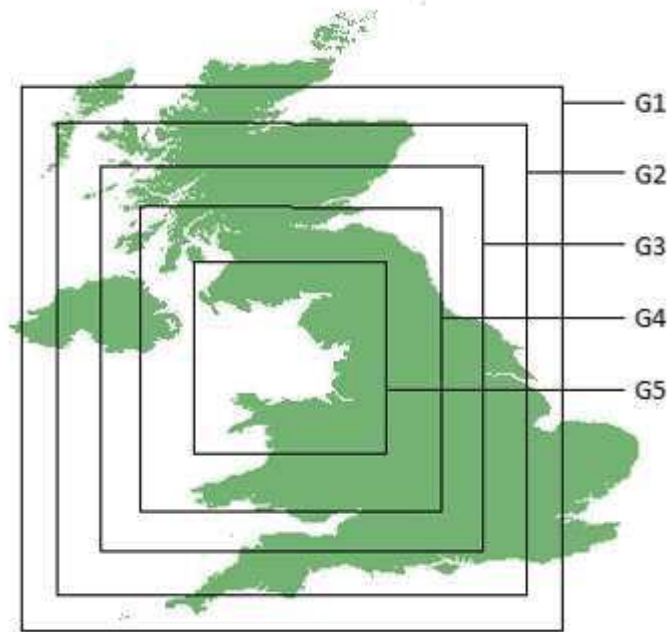


Figure 18. Extent of the  $G_1$  to  $G_5$  sub-regions within the United Kingdom. For simplicity we have represented changing region of interest as a spatial change, however it should be remembered that  $G$  is an ecological entity, not necessarily a spatial one. We can also increase  $G$  by increasing the number of parameters in the model. Normally, as the spatial extent of  $G$  increases we encounter new ecological parameter values not previously studied: this is the increase in  $G$  we consider here.

#### *Environmental variables and virtual species data*

Environmental variables used were annual precipitation (millimetres), elevation (meters above sea level), and annual mean temperature (degrees centigrade\*10), taken from the Worldclim database (Hijmans *et al.* 2005) for the region of the United Kingdom (Figure 19, top row). Five virtual species are then generated in this region, in a similar manner to Elith and Graham 2009. Ecologically, we can consider these as any particular species, epidemiologically they could be thought of as environmental pathogens or vectors. The ecological niche for each of these 5 species responds to these three environmental variables, where the probability of detecting the species at each cell in the region of interest is  $(E*T*P)/3$ .  $E$  is the suitability of the elevation,  $T$  is the suitability of the temperature and  $P$  is the suitability of the precipitation as shown in Figure 19, middle row. This is the true realised ecological niche, normally unknown for real species. The niche is projected into space (Figure 19, bottom row) and a distribution of individual members of the species is generated using the *rbinom* function in *R* (R Core Team 2013) where the probability of a true presence is determined by the suitability of the cell, so the species is more common in suitable habitat. A sample of this presence data is then selected at random, which assumes all ecologies have received equivalent sampling effort.

### *Model construction*

We then use maximum entropy modelling to describe the niche from these samples. Maximum entropy has been discussed in depth previously (Phillips, Dudík and Schapire 2004). In brief, the difference between the distribution of the environmental variables and the species distribution is minimised using the Kullback–Leibler divergence (Phillips, Dudík and Schapire 2004). This effectively assumes the species has the most energetically efficient distribution, which we note is an assumption that is not always ecologically valid. For example, this assumption can be violated for vectors in situations where sexual selection determines distribution (there will be large temporary clusters of the species for breeding), and care should be taken to consider a vector or pathogen in ecology before it is modelled. These maximum entropy models are constructed using the software Maxent 3.3.3k (Phillips, Dudik and Schapire, 2007, Maxent Software, ver. 3.3.3k), which can fit relationships with linear, hinge, threshold, and quadratic functions. We modelled the 5 different virtual species (V) with 10 different sample sizes (u) and 5 different areas of interest (G), with each Maxent model replicated 100 times, resulting in 25,000 models.

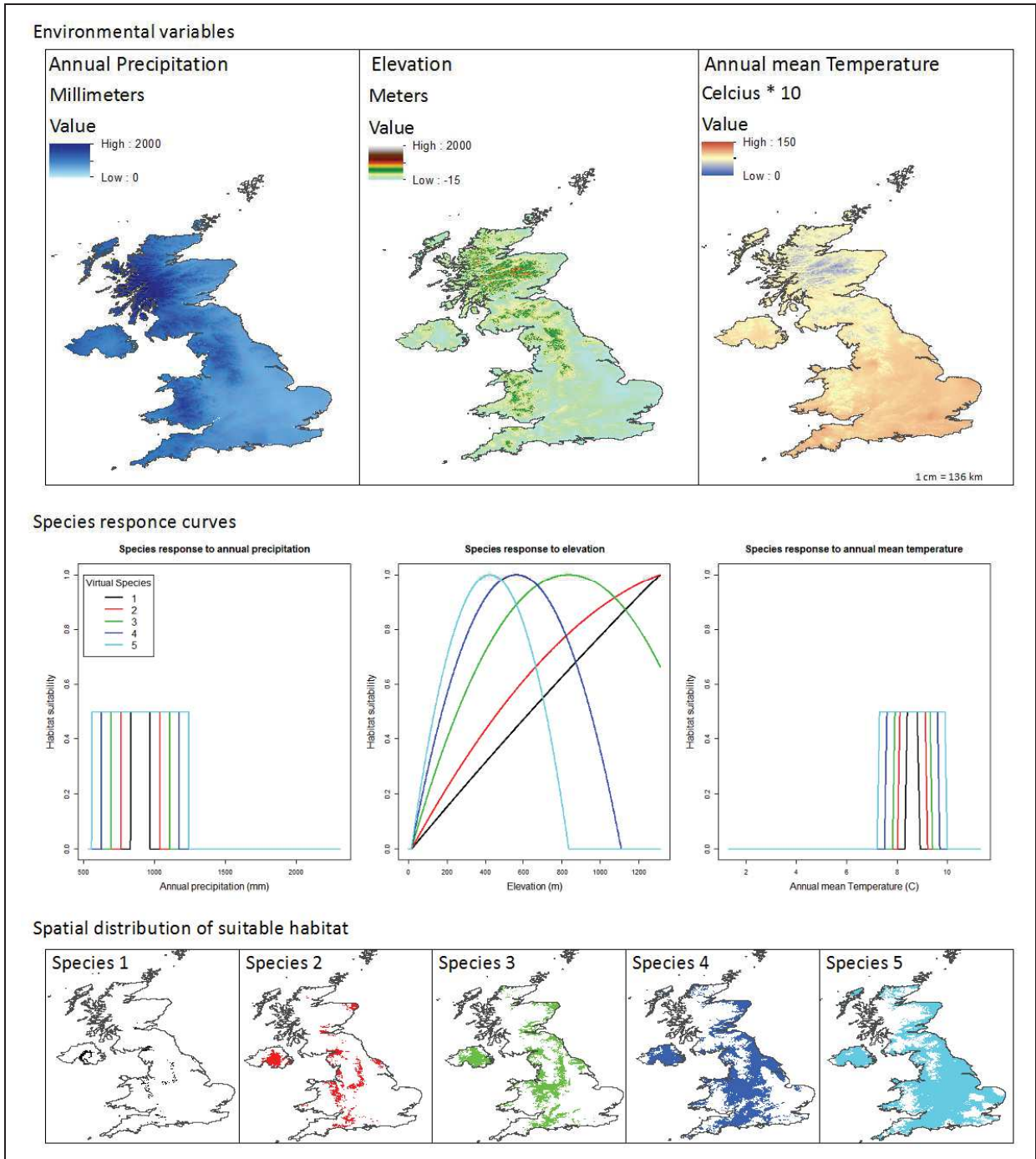


Figure 19. Ecological niches of five virtual species. Annual precipitation, elevation, and annual mean temperature, top row, are used to define the species niches. Each species responds to these variables differently, as shown in the species response curves in the middle row. The different responses result in a different ecological niche, projected into space in the bottom row. A true species distribution is then generated, where the probability of an individual being present in a given cell is proportional to the suitability of that cell, and a sample is then collected from this true distribution.



### *Evaluating model performance*

We aim to explore how changing the sample size, species niche, and region of interest effect the model performance, and to do this we used the AUC as an indicator of performance. The AUC, or area under the curve of the receiver operating characteristic, is a commonly used method of model evaluation, normally derived from presence and absence data, and has been discussed and criticized previously (Lobo *et al* 2008).

AUC is derived from a confusion matrix. Model predictions are classified as correct or incorrect according to the original presence and absence data. An AUC of 1 indicates no false positive or false negative classification errors, a value of 0.5 indicates the model is no better than a random classification, and lower than 0.5 indicates more incorrect classifications than correct. However, often absence data is unavailable due to difficulty in collection or other methodological limitations. This can often happen in infectious disease surveillance where health authorities focus on the presence of a given vector of interest, such as *Aedes albopictus* in southern France, and its absence is considered irrelevant in data collection. In this case we use presence-only modelling methods such as Maxent. However, this raises the problem of model evaluation, as we cannot construct a confusion matrix without absences. To solve this, Maxent normally uses randomly generated pseudo-absence in lieu of real absence data. The use of these pseudo-absences represents the hypothesis that a species distribution should be non-random; however, this hypothesis is not always ecologically valid, as discussed in the conclusion. Maxent normally generates two AUC scores; training AUC, generated with the data used in model training, and test AUC which, herein, is generated with a separate 20% subset of the training data. We also generate a third metric, the true AUC, using the virtual presence data and the real absence data, as opposed to pseudo-absence data which maxent normally uses. This allows us to describe the real performance of the model. The true AUC allows us to explore the ability of test and train AUC to indicate model performance, if test and train AUC are perfect indexes of model performance, they should be correlated to true AUC.

## Results

We found that model AUC increases as sample size increases, increases as niche size decreases (Figure 20), and increases as we increase the region of interest (Figure 21). However, model AUC and true AUC are not always correlated, and the relationship between them varies non-randomly. Model AUC is more accurate (closer to true AUC) for large sample sizes and small niches (Figure 20), which is to be expected, as large sample sizes contain more information, and smaller niches require less information to explain, resulting in a true increase in performance.

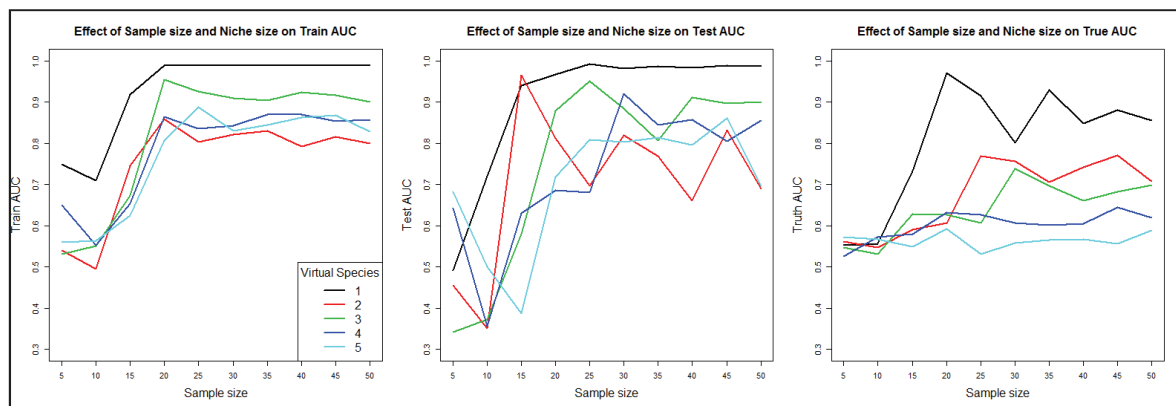


Figure 20. Effect of species niche size ( $V$ ) and sample size ( $u$ ) on model AUC and true AUC. Region of interest is kept constant at  $G_1$ . Colour indicates virtual species niche size, as in Figure 5. Train AUC is the average training AUC of 100 replicates of Maxent models, test AUC is the average AUC generated from 20% subsample validation in these Maxent models and true AUC is the AUC for the same replicate using known presence and absence data, rather than pseudo-absence. Model AUC is always higher for species with smaller niches, but is always over-predicted - true AUC is always lower than model AUC. The difference between model AUC and true AUC is greater for species with larger niches.

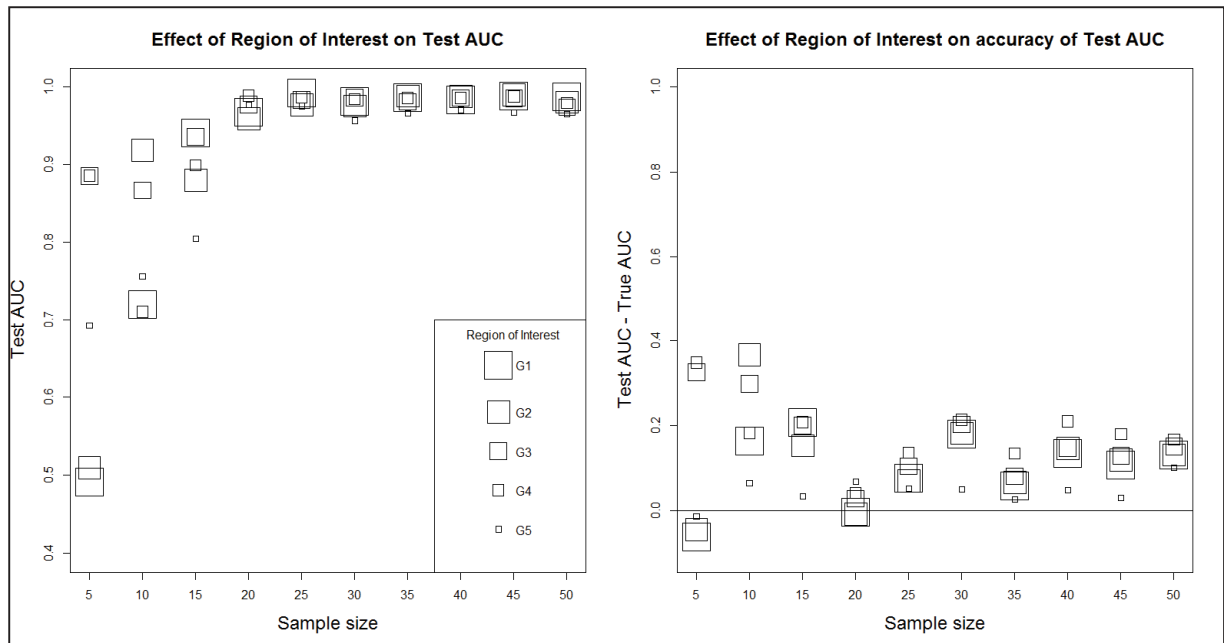


Figure 21. Effect of region of interest (**G**) and sample size ( $u$ ) on model AUC and true AUC. For clarity only virtual species 1 is shown. The left hand figure demonstrates effect of region of interest (as defined in Figure 4) on test AUC. Train AUC shows the same relationship. Larger regions of interest tend to have higher AUC scores. The right hand figure demonstrates the accuracy of these scores, as the difference between test AUC and true AUC. The horizontal black line is zero difference, if the Test AUC is accurate then there should be no difference between true AUC and test AUC. Points above this line indicate over estimated performance, below indicate underestimated performance. Though test AUC is often higher for larger regions of interest, increasing the region of interest tends to result in less accurate AUC scores. The increase is artificial.

The region of interest also has a large effect. For a constant niche and sample size the model AUC will increase as **G** increases (Figure 21, left hand), however there is no corresponding increase in true AUC (Figure 21, right hand). Models constructed on larger regions of interest have a higher AUC, but are less accurate and this increased AUC does not represent a true improvement in performance. This is a result of the use of pseudo-absences in presence-only AUCs and their effect on the confusion matrix. Though presence-only niche models often report good performance for large regions of interest, this is likely an over-estimated AUC.

In sum, our results show that the sample size needed for an adequate model depends on the species and ecological context of the study, and AUC should be used as an index of model performance with caution. Models constructed on different regions of interest cannot be easily compared with the AUC, for example a niche model of the distribution of an insect vector in one region can be inappropriate for comparison with another model, and valid comparison requires equivalent sample size and region of interest.

## Discussion

How many samples do we need to understand the spatial distribution of a pathogen? First ask; where are we looking? When we need to describe the spatial distribution of an emerging disease, and identify areas at risk to future invasion, we need to think of them as species in an ecological context. We show that the sample size (i.e. number of ecologically unique sample sites) needed for ENM changes depending on the size of the pathogen niche ( $\mathbf{B} \cap \mathbf{A}$ , Figure 15), which in turn changes depending on the set of ecological variables in the area under study ( $\mathbf{G}_1$  and  $\mathbf{G}_2$ , Figure 17). There is no fixed sample size needed for niche modelling, no single magic number, rather the amount of ecologically unique samples required changes depending on the ecological region studied. Presence-only models of pathogens, vectors and emerging diseases are often forced to use pseudo-absence data for model evaluation, and this results in an artificially inflated model AUC, particularly for large  $\mathbf{G}$  (Figures 6 and 7). This inflation in improvement is influenced by the size of the species' ecological niche; as niche size increases the maximum achievable model AUC decreases (Wiley *et al* 2003). This change in model AUC is not trivial; as such models can be highly misleading and have important consequences for disease surveillance and management of human health.

Our study has been limited to the use of Maxent, and the degree to which performance is over-estimated is known to vary for different modelling techniques (Elith and Graham, 2009) so caution should be employed in applying our results to other presence-only methods, such as ENFA (Hirzel *et al* 2002). We also cannot comment on sampling bias, as we have assumed that all ecological regions received equivalent sampling efforts (Barry and Elith 2006, Varela *et al* 2009). We examined the relationship between niche size and model performance; however, as we increased the volume of the ecological niche of our virtual species we also increased their spatial distribution (Figure 19). Future work may consider separately examining the effect increasing niche volume without changing the spatial distribution, as the spatial area occupied by the species can influence model performance. The maximum AUC achievable by the model is exactly  $1-(a/2)$ , where  $a$  is the proportion of the spatial area of study covered by the species (Wiley *et al* 2003). As the region of interest increases the niche size can increase, but as  $\mathbf{G}$  continues to increase  $\mathbf{B} \cap \mathbf{A}$  will stop increasing, and begin to cover a smaller proportion of the area of study ( $a$ ). Effectively, as we increase the region of interest, we will artificially increase the model AUC at the same time.

These results are important in the planning for future field work on emerging diseases; the sample size required for ENM will change depending on the goal of the study and the region of interest selected. If initial exploratory analysis can be conducted across this region of interest, such as a particular health district, the volume of ecological variable combinations (the size of  $\mathbf{G}$ ) to be explored can be quantified and the number of sample sites selected can be based on this, allowing

interpolation to every ecological combination in the region of interest. Researchers often use historical data to form a baseline against which to identify an epidemic, using data from previous studies, archaeological explorations, or museum collections, where collection of additional data is often impossible. In this case, a potential tool to select **G** is limit the study to the region appropriate for the available sample size, avoiding extrapolation beyond the ecological range of the sites studied. We recommend against using political boundaries, as they rarely have an ecological justification, but a potential tool in this approach could be multivariate environmental similarity surfaces (MESS maps, Elith, Kearney and Phillips 2010).

Currently, in an era of large scale ecological disturbance and climate change, the distributions of many environmentally acquired and zoonotic disease agents are rapidly changing. Future cross-disciplinary research on their current and potential spatial distributions requires an ecological approach; we need to appropriately delineate the region of study based on ecological assumptions about the species, and consider the pathogen from a less anthropocentric view; but rather as a species in ecology.

## References

- Anderson R, *et al.* (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2): 129-151.
- Ayala D, *et al.* "Habitat suitability and ecological niche profile of major malaria vectors in Cameroon." *Malar J* 8.1 (2009): 307.
- Barve N, *et al.* (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, 222(11), 1810-1819.
- Daszak P, (2013). Interdisciplinary approaches to understanding disease emergence: The past, present, and future drivers of Nipah virus emergence. *Proceedings of the National Academy of Sciences*, 110 (Supplement 1), 3681-3688.
- Dudík, M., *et al.*, (2007). Maximum Entropy Density Estimation with Generalized Regularization and an Application to Species Distribution Modeling. *Journal of Machine Learning Research*, 8(6).
- Elith J, *et al.* (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, 17(1), 43-57.
- Elith J, and Graham CH, (2009). Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography*, 32(1), 66-77.
- Elith J, Kearney M, and Phillips S, (2010). The art of modelling range-shifting species. *Methods in ecology and evolution*, 1(4), 330-342.
- Hay S, *et al.*, (2013). Big data opportunities for global infectious disease surveillance. *PLoS medicine*, 10(4), e1001413.
- Hernandez P, *et al.*, (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, 29(5), 773-785.
- Hijmans R, *et al.* (2005). Very high resolution interpolated climate surfaces for global land areas. *International journal of climatology*, 25(15), 1965-1978.
- Hirzel A, *et al.*, (2002). Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data?. *Ecology*, 83(7), 2027-2036.
- Barry S, and Elith J, (2006). Error and uncertainty in habitat models. *Journal of Applied Ecology*, 43(3), 413-423.
- Hutchinson GE, (1957). "Concluding remarks" (PDF). Cold Spring Harbor Symposia on Quantitative Biology 22 (2): 415-427.
- Lobo JM, *et al.*, (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17(2), 145-151.
- Meentemeyer R, *et al.*, (2004). Mapping the risk of establishment and spread of sudden oak death in California. *Forest Ecology and Management*, 200(1), 195-214.

Mullins J, *et al.* (2011). Ecological Niche Modelling of the Bacillus anthracis A1. a sub-lineage in Kazakhstan. *BMC ecology*, 11(1), 32.

Matthews, J. N., et al. "Analysis of serial measurements in medical research." *BMJ: British Medical Journal* 300.6719 (1990): 230.

Saupe EE., *et al.*, (2012). Variation in niche and distribution model performance: the need for a priori assessment of key causal factors. *Ecological Modelling*, 237, 11-22.

Soberón J, and Nakamura M, (2009). Niches and distributional areas: concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences*, 106(Supplement 2), 19644-19650.

Phillips, S. J., Dudík, M., and Schapire, RE (2004). A maximum entropy approach to species distribution modeling. *Proceedings of the twenty-first international conference on Machine learning* (p. 83). ACM.

Phillips SJ, Anderson RP, and Schapire, RE (2006). Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3), 231-259.

Peterson AT, and Robins CR, (2003). Using Ecological-Niche Modeling to Predict Barred Owl Invasions with Implications for Spotted Owl Conservation. *Conservation Biology*, 17(4), 1161-1165.

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

Tran A, et al. "Using remote sensing to map larval and adult populations of Anopheles hyrcanus (Diptera: Culicidae) a potential malaria vector in Southern France." *International Journal of Health Geographics* 7.1 (2008): 9.

Varela S, *et al.*, (2009). Is current climatic equilibrium a guarantee for the transferability of distribution model predictions? A case study of the spotted hyena. *Journal of Biogeography*, 36(9), 1645-1655.

VanDerWal J, *et al.*, (2009). Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know?. *ecological modelling*, 220(4), 589-594.

Wiley EO, *et al.* (2003). Niche modeling and geographic range predictions in the marine environment using a machine-learning algorithm. *Oceanography*, 16(3), 120-127.

**Chapter 3: Topography and land cover of watersheds predicts the distribution of the environmental pathogen *Mycobacterium ulcerans* in aquatic insects.**



## Foreword.

The following chapter has been editorially accepted as a paper to the journal of PLOS Neglected Tropical Diseases, titled “Topography and land cover of watersheds predicts the distribution of the environmental pathogen *Mycobacterium ulcerans* in aquatic insects”, by Kevin Carolan, Andres Garchitorena, Gabriel Garcia-Pena, Aaron Morris, Jordi Landier, Arnaud Fontanet, Philippe Le Gall, Gaëtan Texier, Laurent Marsollier, Rodolphe E. Gozlan, Sara Eyangoh, Danny Lo Seen, and Jean-Francois Guégan.

This chapter uses the data collected in Cameroon by Andres Garchitorena, with land cover maps I constructed, to make a model of the niche of *M. ulcerans* and test it in French Guiana with data from Aaron Morris. As discussed within, the model in Cameroon does not predict the bacteria in French Guiana well. Speculating on potential reasons for this, we suspect that either the bacterium may have evolved to occupy a new niche in French Guiana, or it may be that the ecological context of South America is too different to be comparable to Africa, however this is strictly speculative.

The generation of the landcover maps used herein involved a large investment of the time taken to complete this thesis. The process of landcover map generation is often seen as a simple procedure, however there are several nuances that make this time consuming. A large amount of time was spent learning to use several different softwares; ArcMap, eCognition, ENVI. Several satellite images were then used, SPOT, Rapideye and Landsat, to generate preliminary maps, which were ground truthed in a three week field mission to Cameroon in November 2012 (grateful thanks are extended to Sara Eyangoh and Gaëtan Texier of the Centre Pasteur du Cameroun for support on this field mission). Following this mission, the maps were redrafted to improve accuracy. The experience in the field was extremely useful in generating several hypotheses about the role of watersheds, as real experience of the shape of the terrain helped to develop thought about how the aquatic community that influences the bacteria would be distributed.

This research involved extensive collaboration with Andrés Garchitorena, who was supported by a PhD studentship from the École des hautes études en santé publique (EHESP), and Aaron Morris who was supported by a PhD studentship from Bournemouth University. Gabriel García-Peña was supported by a postdoctoral fellowship from the French Fondation pour la Recherche sur la Biodiversité (CESAB research centre, research programme BIODIS). I would also like to thank Annelise Tran of CIRAD and Benjamin Roche of IRD for invaluable discussions and insights on the manuscript, the ISIS Spot programme for support in acquiring SPOT images, and Hervé Chevillotte (IRD Cameroon), for environmental data from the IFORA project (ANR-Biodiv grant IFORA).

## Author Summary

Many pathogens persist in the environment, and understanding their distribution can assist in control, allowing us to identify areas of risk to local human populations. Herein, we describe the distribution of a particular environmental pathogen, *Mycobacterium ulcerans*, describing the landscape conditions correlated with its presence in local biota, and mapping the distribution of these habitats in a region of Cameroon, Africa. We identify the importance of the watershed as a factor determining the distribution of the bacterium, where conditions removed from the sample site can influence the abundance of the bacterium in downstream sites. We also discuss sensitivity of these models to extrapolation, working well in the African region and failing in another region, in South America.

## Background

Knowledge of the spatial distribution of an environmentally persistent pathogen is often key in creation of environmental hazard maps for disease control. Yet, despite the importance of this spatial information, only 4% of such pathogens have been mapped (Hay *et al.* 2013). The reason for this gap in our knowledge is practical. It is often difficult to produce large maps of the distribution of these microbial pathogens as they are difficult to detect in nature. A solution to this is to describe the distribution of the pathogens suitable habitat. For example, an environmentally persistent pathogenic bacterium may have a certain *pH* range within which it can survive, a specific range of microaerobic oxygen concentrations (Palomino *et al.* 1998), and survive preferentially on certain algae (Marsollier *et al.* 2004A). In cases where we have a suitable range of *pH*, a suitable range of oxygen, and suitable algae, we expect to find the bacterium. Herein, this suitable range of microhabitat is termed the ecological niche of the species. Every species in nature, including vectors such as mosquitoes, and pathogens such as *Plasmodium* protozoans, has a unique ecological niche (Hutchinson 1957, Soberon 2005).

Knowledge of the distribution of suitable habitats would allow us to predict the expected distribution of the pathogen. This approach has been successfully applied to the vectors of diseases such as malaria, plague and dengue (Tran *et al.* 2008, Ayala *et al.* 2009, Ari *et al.* 2012), but it is rarely applied to environmentally persistent pathogenic microbes. The range of suitable habitat is, practically, much easier to describe for insect vectors than for microbes. For example, the suitable habitat of mosquitos is driven by factors such as rainfall, which is much easier to describe on a large scale. To describe *pH* in the environment we must visit each site and use a probe at each location.

This quickly becomes expensive and time consuming when we consider multiple variables, or if we wish to describe the distribution of a pathogen over large extents.

We hypothesised that these microhabitat variables could be indirectly inferred from large scale macroecological patterns. The distribution of swamp and forested environment, the shape and structure of the landscape, should predict the distribution of these microhabitats. For example, while the suitable habitat of a bacterium may be driven by the suitable combination of *pH*, oxygen, and algae, and other factors, the distribution of these conditions is in turn driven by the landscape. For example, the *pH* and oxygen content of water in swamps is lower, on average, than of water in savannahs. We can use the landscape, which is more easily described, as a proxy to describe the spatial distribution of this suitable microhabitat. Though this approach is limited in lacking a physiological understanding of direct influences on the pathogen, it has the great benefit of inferring the potential distribution of the pathogen, opening new opportunities to disease control.

We undertook ecological niche modelling of *Mycobacterium ulcerans*, an environmentally acquired pathogenic bacterium, and causative agent of Buruli ulcer. The ecological niche refers to this range of conditions within which a species can survive and maintain a population. We infer that, if a species has a large population, it presumably is able to maintain that population, and is in a suitable environment. By understanding the environmental parameters that describe population size, we can predict the distribution of the pathogen. Maps of the distribution of pathogens are often a key step in control of disease, producing environmental hazard maps.

The pathogen of our study, *Mycobacterium ulcerans*, infects up to 10,000 people per year in more than 30 countries around the world (Johnson *et al.* 2005, WHO 2008). Infection leads to the Buruli ulcer, an emerging neglected tropical disease (WHO 2008) which results in a necrotizing infection of the skin and can lead to crippling deformity (Johnson *et al.* 2005). The transmission route of *M. ulcerans* remains unknown, and though several competing hypotheses exist (Marsollier *et al.* 2002, Benbow *et al.* 2008) our work herein does not address transmission, but focuses on the distribution of the pathogen.

Identification of the landscape variants that indicate suitable habitat for this particular pathogen has proven remarkably difficult, despite decades of research (see Merritt *et al.* 2010 for a review). Previous research on *M. ulcerans* has found several apparently contradictory facts about the bacterium, making it difficult to establish a generalised picture of its ecology. In 2007 the genome of *M. ulcerans* was sequenced, and analysis revealed extensive evidence for reductive evolution, with massive gene loss. *M. ulcerans* evolved from *M. marinum*, and appears to have undergone a

bottleneck event in the process, losing many of the genes *M. marinum* uses to sustain itself in free living environments, apparently now favouring protected environments with low sunlight (Stinear *et al.* 2007). This is suggestive of a highly specialised ecological niche, implying that the bacterium cannot survive in a large range of environmental conditions. Detection of the bacterium in the environment is normally via PCR; *M. ulcerans* is very slow growing and extremely difficult to culture from the wild (Portaels *et al.* 2008), and most attempts at culture result in *M. ulcerans* being overgrown by other bacteria which are ubiquitous in the environment.

However, the implication that the microbe is a specialist has been (apparently) contradicted by recent detection of the bacterium in the environment. *M. ulcerans* DNA has been detected in a bewildering variety of environmental samples, including aquatic insects, biofilms, crustaceans, detritus, fish, frogs, possums and various small mammals, soil, snails, water and worms (Portaels *et al.* 1999, Stinear *et al.* 2000, Portaels *et al.* 2001, Eddyani *et al.* 2004, Marsollier *et al.* 2004A, Trott *et al.* 2004, Kotlowski *et al.* 2004, Johnson *et al.* 2007, Fyfe *et al.* 2007, Williamson *et al.* 2008, Portaels *et al.* 2008, Fyfe *et al.* 2010, Roche *et al.* 2013, Carson *et al.* 2014, Morris *et al.* 2014 A, Garchitorena *et al.* 2014). This large range of suitable conditions is odd, in light of the bacterium's apparent status as a specialist with a small niche.

The many different species that *M. ulcerans* infects in the local community may become infected due to differences in their feeding habits, position in the trophic web, or relative abundance (Marsollier *et al.* 2004B, Mosi *et al.* 2008, Merritt *et al.* 2010). Herein, we use samples of the five dominant Orders of the aquatic insect community, which have been tested for *M. ulcerans* positivity rates, and correlate changes in *M. ulcerans* positivity in these 5 Orders to changes in the environmental conditions of land cover and topography. These 5 Orders may not be the primary habitat of *M. ulcerans* in the wild, as the full biotic extent of *M. ulcerans* distribution is still unknown, but they are commonly found to be persistently infected and appear to be important hosts (Marion *et al.* 2010). Previous work has found that *M. ulcerans* abundance does respond to water body type, being more commonly detected in swamps (still lentic systems) than rivers (flowing lotic systems) in Ghana (Benbow *et al.* 2013, McIntosh *et al.* 2014). The pathogen is associated with lowland, flat, swampy areas in contact with stagnant water (Wagner *et al.* 2008 A), is known to have complex seasonal dynamics (Marion *et al.* 2010), and appears to be present at low levels throughout the entire local biotic community along the year (Garchitorena *et al.* 2014). The distribution of the disease may also inform us on the distribution of the pathogen; the distribution of Buruli ulcer is known to be more spatially restricted than the distribution of *M. ulcerans* (Williamson *et al.* 2012), and is known to respond to low elevation, forested land cover, and previous rainfall (van Ravensway

*et al.* 2012, Morris *et al.* 2014 B), which would suggest that perhaps these factors are also important in the distribution of *M. ulcerans*. Taken together, these facts suggested that changes in the biotic distribution of the pathogen could be mapped using landscape variables. Often, sampling of river systems results in the unexpected presence of *M. ulcerans*; if factors at the larger watershed scale add substantial information on the distribution of *M. ulcerans* a description of the upstream region of the river may help to explain this unexpected presence. We describe the condition of the landscape using land cover, such as forest and savannah, and topography, such as elevation and slope. These landscape scale factors are expected to indirectly influence *M. ulcerans* abundance via their influence on the microhabitat the bacterium inhabits, for example affecting the *pH*, dissolved oxygen content, and composition of the aquatic insect community, which are known to influence *M. ulcerans* distribution (Benbow *et al.* 2008, Garchitorena *et al.* 2014).

To address our questions we describe landscape variables correlated to the presence of the bacterium in aquatic macroinvertebrates in Cameroon, Central Africa. We then test our model against data collected in French Guiana to explore the generalizability of our findings. This will contribute to an understanding of the spatial distribution of this environmental pathogen, and further our ability to control Buruli ulcer disease.

## **Methodology**

A model was constructed on the dataset from Akonolinga, Cameroon, and predicted into French Guiana, South America. This enabled us to describe the niche of *M. ulcerans*, and examine how well these models transferred to other areas.

### *Study sites, sampling methodology and response variable*

The Cameroon dataset is a subset of that published in Garchitorena *et al.* (2014), which comprises 16 sites in Akonolinga, sampled every month for 12 months (Figure 22). Identical methods were carried out by the same investigators for all sites throughout the study. In brief, at each site, 4 locations were chosen in areas of slow water flow and among the dominant aquatic vegetation and at each location, 5 sweeps with a dip net within a surface of 1m<sup>2</sup> were done to sample the aquatic community. Aquatic organisms were classified down to the Family level whenever possible and stored separately in 70% ethanol. Individuals belonging to the same taxonomic group were pooled together for detection of *M. ulcerans* DNA by quantitative PCR. Among these, the 5 most abundant Orders (Diptera, Hemiptera, Coleoptera, Odonata and Ephemeroptera) were consistently analysed for all sites and months. Pooled individuals were all ground together and homogenized and DNA

from tissue homogenates was purified using QIAquick 96 PCR Purification Kit (QIAGEN). Finally, amplification and detection of MU DNA were performed through quantitative PCR by targeting the ketoreductase B domain (KR) of the mycolactone polyketide synthase and IS2404 sequence from MU genome. This resulted in 5 analyzed samples (each Order) per month, per site, which we use to infer *M.ulcerans* presence or absence. Summary statistics are described in Table 1. Sampling effort varied from month to month, as is discussed in Garchitorena *et al.* (2014), however we have used a subset of that data in order to gain the most consistent representation of the biotic community possible.

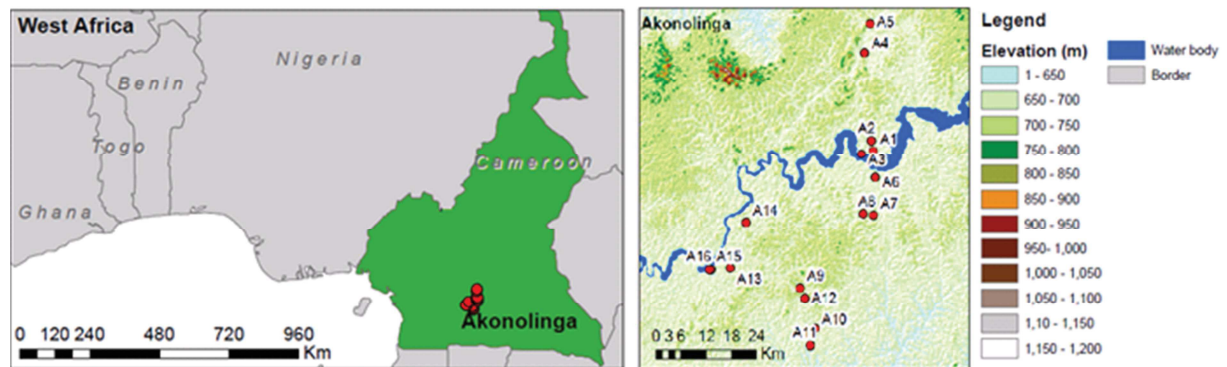


Figure 22. Location of sample sites in Cameroon, as in Garchitorena *et al.* 2014. Within Cameroon, Akonolinga is almost entirely rainforest. This region is dominated by the Nyong river and has fewer highland areas. Red dots are sample sites in Akonolinga.

A data set following the same methodology was independently collected in French Guiana, South America (Morris *et al.* 2014). DNA extraction was carried out with the same two primer pairs and methodology as above. In French Guiana eighteen sites were sampled twice during the wet season, which lasts from December to July. The entire biotic community was sampled, and for consistency the same 5 taxonomic Orders as in Akonolinga (Table 2) were compared.

Table 1. *M. ulcerans* distribution at sample sites in Akonolinga, Cameroon. 16 sites were sampled for 12 months, sampling from different types of water bodies. The dominant members of the aquatic biota were Diptera, Hemiptera, Coleoptera, Odonata and Ephemeroptera. These made up the majority of the community in both seasons, the percentage of the biotic sampled community composed of these five groups is reported as Relative abundance of the 5 Orders in Table 1. These communities were normally positive of *M. ulcerans*, the percentage of positive samples (number of positive samples/total samples for the 5 Orders from that site in that season) describes the PCR positive samples of the 5 Orders. This table is a summary of a subset of the data presented in Garchitorena *et al.* (2014).

Site Code	Latitude	Longitude	Type of water body	Wet season		Dry season	
				Relative abundance of 5 Orders (%)	PCR positive samples of the 5 Orders	Relative abundance of 5 Orders (%)	PCR positive samples of the 5 Orders
A1	N3°46.806	E12°16.133	Swamp	93.69	6.98% (3/43)	97.39	20% (4/20)
A2	N3°47.083	E12°15.383	Swamp	94.69	12.36% (11/89)	94.53	8.11% (3/37)
A3	N3°46.316	E12°14.440	Stream	90.82	2.56% (1/39)	90.41	0% (0/20)
A4	N3°58.464	E12°14.796	River	55.12	5.13% (2/39)	32.17	5.26% (1/19)
A5	N4°02.255	E12°15.620	Stream	83.64	10% (4/40)	75.30	5.26% (1/19)
A6	N3°43.483	E12°16.466	Swamp	90.31	17.24% (15/87)	85.85	2.94% (1/34)
A7	N3°38.889	E12°15.986	River	79.52	10.26% (4/39)	79.13	0% (0/13)
A8	N3°38.980	E12°14.696	River	64.86	9.09% (4/44)	73.89	7.69% (1/13)
A9	N3°29.912	E12°06.425	Swamp	82.40	4.4% (4/91)	91.67	2.7% (1/37)
A10	N3°29.912	E12°06.425	Swamp	93.62	7.5%(3/40)	89.90	0% (0/20)
A11	N3°23.271	E12°07.870	River	50.12	4.65% (4/86)	59.56	8.82% (3/34)
A12	N3°28.788	E12°07.255	River	52.87	2.44% (1/41)	66.19	15% (3/20)
A13	N3°32.322	E11°57.643	Swamp	81.92	2.22% (2/90)	79.39	8.11% (3/37)
A14	N3°38.032	E11°59.695	River	64.77	15.38% (6/39)	63.37	20% (4/20)
A15	N3°32.288	E11°55.239	Flooded area	94.33	11.11% (4/36)	97.54	0% (0/15)
A16	N3°32.276	E11°55.181	Stream	89.85	2.86% (1/35)	93.40	0% (0/15)

Table 2. *M. ulcerans* distribution at sample sites in French Guiana, South America. 18 sites were sampled in the wet season. The dominant members of the aquatic biota were Diptera, Hemiptera, Coleoptera, Odonata and Ephemeroptera, as in Akonolinga. These made up the majority of the community, the percentage of the biotic sampled community composed of these five groups is reported as Relative abundance of the 5 Orders. These communities were normally positive of *M. ulcerans*, the percentage of positive samples (number of positive samples/total samples for the 5 Orders from that site in that season) describes the PCR positive samples of the 5 Orders. This table is a summary of a subset of the data presented in Morris *et al.* (2014 A).

Site Code	Latitude	Longitude	Wet season	
			Relative abundance of 5 Orders (%)	PCR positive samples of the 5 Orders,% (positive/samples)
FG10	N4 44.170	W-52 19.618	58.62	44.12% (15/34)
FG11	N4 50.284	W-52 21.195	52.54	32.26% (10/31)
FG19	N5 17.773	W-53 03.085	62.50	10.00% (1/10)
FG2	N5 37.888	W-53 42.433	70.83	11.76% (6/51)
FG23	N5 21.724	W-53 2.0200	29.27	16.67% (2/12)
FG28	N5 36.328	W-53 49.660	56.60	20.00% (6/30)
FG34	N4 50.068	W-52 18.126	85.00	32.35% (11/34)
FG38	N5 23.646	W-52 59.521	73.74	13.70% (10/73)
FG41	N5 25.725	W-53 05.326	41.07	8.70% (2/23)
FG43	N5 22.632	W-52 57.232	75.47	2.50% (1/40)
FG44	N4 20.052	W-52 09.148	25.42	0.00% (0/15)
FG45	N4 18.025	W-52 07.397	61.95	2.86% (2/70)
FG46	N5 02.121	W-52 30.989	74.24	2.04% (1/49)
FG47	N4 55.744	W-52 24.229	65.00	0.00% (0/26)
FG48	N4 51.616	W-52 16.518	16.67	0.00% (0/1)
FG49	N5 39.996	W-53 46.794	36.54	0.00% (0/19)
FG53	N5 36.136	W-53 50.182	67.86	0.00% (0/57)
FG7	N4 51.648	W-52 15.405	29.41	0.00% (0/10)

#### *Seasonal effects on M. ulcerans distribution*

*M. ulcerans* has previously been found to respond to variables that are influenced by rainfall (Wagner *et al.* 2008 A, Morris *et al.* 2014 B). To explore differences in the seasonal distribution of the bacterium, the wet season months and the dry season months were analysed separately. In Cameroon wet season months are April, May, June, August, September and October. The dry season is January, February, March, July, November and December. For each site, the proportion of positive samples at a site in a season was determined by summing the number of positive samples in that season, then dividing by the total number of samples sampled in that season (which is 5 multiplied by the number of sampled months). This resulted in two response variables,  $Y_{\text{wet}}$  and  $Y_{\text{dry}}$ , which we use to describe the proportion of *M. ulcerans* positive samples in the 5 dominant insect Orders in the wet and dry seasons respectively. This resulted in a general, standardised view of the



mycobacterium distribution in both the dry and wet seasons. The habitat suitability is determined by the proportion of samples of the biotic community that are *M. ulcerans* positive.

#### *Land cover and topography*

Land cover in Akonolinga was described using several multispectral satellite images; SPOT 2.5 meter resolution images (references: 50833380811220923092V0 and 50833371012210937422V0), and a Landsat image (reference L72186056\_05620021107). The study area was categorised into the following classes; Agriculture, Forest, Flood plain, Road, Savannah, Swamp and Urban (Table 3). Classification was conducted in the Object Orientated Image Analysis software eCognition (eCognition Trimble Navigation Ltd). The resulting maps were validated and corrected where needed following onsite visits in November 2012. Topography was described using the Shuttle Radar Topography Mission (SRTM) digital elevation model (Jarvis *et al.* 2008), which has a spatial resolution of 90 meters. All topographical variables were derived using the Spatial Analyst extension of the software ArcMap 10.1 (ESRI 2011). For each site we described the mean, standard deviation, minimum, maximum and variety of elevation, in meters above sea level, using SRTM (Table 3). From the SRTM we calculated the mean, standard deviation, minimum, maximum and variety of the topological slope, in degrees. Flow accumulation is the accumulated number of upstream cells flowing into a point, and ecologically represents the topographical potential for water to accumulate. We derived the mean, standard deviation, maximum and variety of the flow accumulation. We also calculated mean, standard deviation, maximum depth, variety, and proportion of buffer surface area covered by basins. Basins are depressions in the landscape where water is expected to accumulate and, potentially, stagnate, and were detected using the Fill function in Spatial Analyst extension in Arc Map. Stream order indicates the distance from the source of the river, and is a simple index of the type of stream (1<sup>st</sup> order being small streams, larger orders being big rivers). Proportion of 1<sup>st</sup> to 8<sup>th</sup> order streams, defined by Strahler method (Strahler 1957), was recorded in each buffer. Finally, wetness index is the topographic potential for water to accumulate. It was derived from the flow accumulation and the slope, according to the Equation 1, where WI is the wetness index (Beven and Kirkby, 1979), FA is flow accumulation and S is the topographic slope in degrees. We derived the mean, standard deviation, maximum, and variety of wetness index values, and the proportion of buffer surface area covered by wetness index values which are positive (relatively wet areas) and negative (relatively dry areas).

$$WI = \ln \left( \frac{FA}{\tan(S)} \right)$$

Equation 1.

*Importance of local effects compared to regional effects in M. ulcerans distribution.*

The topography and land cover of the sample sites were described within two different buffers (Figure 23). These buffers corresponded to local and regional conditions. The first buffer was a 5km radius circle around the sample site, which was chosen to represent the local conditions. 5km is, approximately, the flight range of the 5 insect orders sampled (Bowden 1964, Robertson 1976, Lytle 1999, Mukai and Ishii 2007). The insects should be able to move throughout this region, be exposed to *M. ulcerans*, before being captured at the sample site. We describe the land cover and topography within this 5km buffer and correlate the condition of this region to the proportion of *M. ulcerans* positive pools in each season.

The second buffer was defined using the watershed of the sample site (Figure 23). The watershed is the upstream catchment area. In principle, all water within this region, and any detritus floating in the water, will eventually flow through the sample site. Watersheds can vary greatly in size, easily being several kilometres long, and detritus from very distant locations can flow quite large distances. *M. ulcerans* is known to attach to such detritus (Williamson *et al.* 2008). This watershed buffer is created using the Watershed tool in ArcMap10.1, Spatial Analyst extension (ESRI).

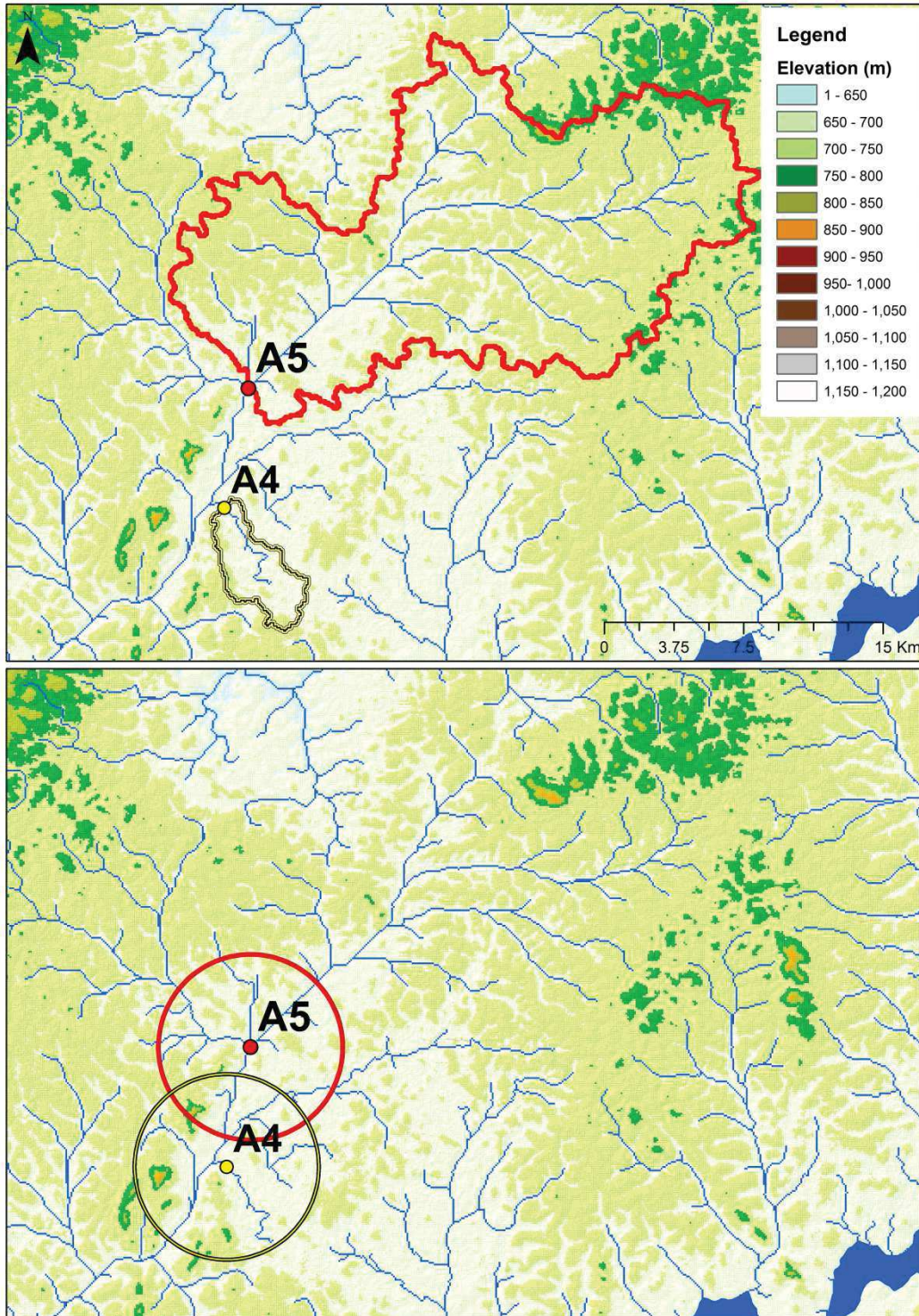


Figure 23. An example of the two buffer types used in this examination, sites A4 and A5 in Akonolinga. This is in the north of Akonolinga, near the village of Emvong. The lower panel is a 5km buffer around the sites, within this region we describe the topography and land cover, and its association with *M. ulcerans* abundance. We compare this to the watershed buffer (upper panel). The watershed is the drainage area for each site, in principle all water that falls within this region will eventually pass through the sample site.

### *Principal component analysis*

The 42 variables estimated to describe the landscape were reduced to permit modelling. Principal component analysis (PCA) was performed on the landscape variables centred at the mean ( $\ln(x) - \ln(x_{\text{mean}})$ ) to summarize the data in the watershed and the 5km buffer. PCAs were performed with the PCA function in the FactoMineR library in R (R Core Team 2014). This generated two PCAs; a PCA of the 42 environmental variables in the watershed buffer,  $\text{PCA}_{\text{ws}}$ , and a PCA of the 42 environmental variables in the 5km buffer,  $\text{PCA}_{5\text{km}}$ . In each PCA we examined the orthogonal axes that explained 95% of the variance in the 42 topography and land cover variables.

Firstly, 9 principal components explained 95% of the variance in the watershed of the sample site ( $\text{PCA}_{\text{ws}}$ ). The magnitude and direction of each correlation is given in the supplementary materials (Tables 3 and 4). We describe  $\text{PCA}_{\text{ws}1}$  as “large watersheds that drain flood plains”, given its strongly positive correlations to watershed surface area and floodplains;  $\text{PCA}_{\text{ws}2}$  as “large watersheds that drain highland agriculture”;  $\text{PCA}_{\text{ws}3}$  as “large watersheds that drain lowland agriculture”;  $\text{PCA}_{\text{ws}4}$  as “small watersheds that drain swamp and forest at flat intermediate elevations”;  $\text{PCA}_{\text{ws}5}$  as “small watersheds that drain highland urban and savannah”;  $\text{PCA}_{\text{ws}6}$  as “small watersheds that drain highland urban and forest”;  $\text{PCA}_{\text{ws}7}$  as “large watersheds that drain lowland forest, savannah and swamp”;  $\text{PCA}_{\text{ws}8}$  as “small watersheds that drain urban and agricultural environments in hilly lowlands”; and  $\text{PCA}_{\text{ws}9}$  as “small watersheds that drain wet swamps in areas that reach from low to high elevations” (Table 3).

Secondly, for the local 5km circular buffer, 6 principal components ( $\text{PCA}_{5\text{km}}$ ) explained 95% of the variance in the data as described in SM2. Translating these to ecologically meaningful terms, we describe  $\text{PCA}_{5\text{km}1}$  as representing “sites surrounded by flat lowland areas with urban, agriculture and the flood plains of large rivers”;  $\text{PCA}_{5\text{km}2}$  as representing “sites surrounded by sloped highland areas with urban, agriculture and small rivers”;  $\text{PCA}_{5\text{km}3}$  as representing “sites surrounded by sloped highland areas with savannah and large swampy rivers”;  $\text{PCA}_{5\text{km}4}$  as representing “sites surrounded by flat lowland areas with savannah and small rivers”;  $\text{PCA}_{5\text{km}5}$  as representing “sites surrounded by flat highlands with urban, agriculture and large rivers”, and  $\text{PCA}_{5\text{km}6}$  as representing “sites surrounded by lowland hills, with small rivers and many small basins, in unforested environment”, (Table 4).

Table 3. Results of Principle Component Analysis for topographical and landcover variables in a watershed buffer. 95% of the variance in the data was described with 9 components, the eigenvalue of each component is given at the bottom of the table. Each component correlates differently to different variables, red highlights negative correlations, blue highlights positive correlations. PCA<sub>ws1</sub> describes large watersheds that drain flood plains and swamps, with few urban and agricultural areas. These are high elevation areas with variable slopes. PCA<sub>ws2</sub> describes large watersheds that drain agriculture at flat highland areas. PCA<sub>ws3</sub> describes large rivers that drain urban and agriculture areas at flat lowlands with, with little forest. PCA<sub>ws4</sub> describes small rivers, with small watersheds that drain forest and swamp areas, without urban areas. These are at intermediate elevations, with flat areas. PCA<sub>ws5</sub> describes small rivers that drain urban and savannah areas, predominantly in higher elevation flat lands. PCA<sub>ws6</sub> corresponds to small low order streams that drain urban and forest (not agriculture) in high elevation slopes. PCA<sub>ws7</sub> is larger watersheds that drain forest, savannah flood plain and swamp, in areas with flat, wet, lowlands. PCA<sub>ws8</sub> represents small watersheds that drain urban & agriculture, flood plain and savannah. These areas are wet lowlands with lots of small hills. PCA<sub>ws9</sub> represents small watersheds that drain wet swamps in areas that reach from low to high elevations.

Covariate	Units	PCA <sub>ws1</sub>	PCA <sub>ws2</sub>	PCA <sub>ws3</sub>	PCA <sub>ws4</sub>	PCA <sub>ws5</sub>	PCA <sub>ws6</sub>	PCA <sub>ws7</sub>	PCA <sub>ws8</sub>	PCA <sub>ws9</sub>
Surface area	m <sup>2</sup>	0.94	0.18	0.19	-0.08	-0.02	-0.04	0.15	-0.03	-0.02
<b>Land cover</b>										
Urban	%	-0.22	-0.12	0.80	-0.10	0.13	0.28	-0.25	0.07	0.05
Agriculture	%	-0.46	0.20	0.68	-0.11	-0.26	-0.02	-0.31	0.06	0.07
Forest	%	-0.37	-0.15	-0.73	0.13	-0.32	0.08	0.01	-0.22	0.04
Roads	%	0.57	-0.06	0.54	-0.28	-0.04	-0.17	-0.28	-0.20	0.26
Savannah	%	-0.04	-0.18	-0.09	0.05	0.84	-0.18	0.23	0.25	0.10
Flood plain	%	0.84	0.10	0.18	-0.12	-0.18	-0.33	0.21	0.14	0.04
Swamp	%	0.18	-0.22	0.02	0.55	-0.21	0.21	0.28	-0.20	0.52
<b>Flow accumulation</b>										
Mean	m <sup>2</sup>	0.97	0.14	0.14	-0.04	-0.02	-0.01	0.11	-0.05	0.03
Sd	-	0.95	0.17	0.18	-0.08	-0.02	-0.04	0.12	-0.05	0.02
Max	m <sup>2</sup>	0.94	0.18	0.19	-0.08	-0.02	-0.04	0.15	-0.03	-0.02
Variety	-	0.96	0.16	0.16	-0.05	-0.03	-0.02	0.15	-0.01	-0.04
<b>Basins</b>										
Mean basin depth	m	0.01	-0.93	0.13	-0.22	0.09	0.06	0.06	-0.20	-0.10
Sd basin depth	-	0.21	0.80	-0.28	0.23	-0.23	-0.07	-0.17	0.28	0.03
Deepest basin	m	-0.90	-0.13	0.20	-0.23	0.20	-0.07	0.02	-0.09	0.10
Variety of basin depth	-	0.93	0.05	-0.19	0.19	-0.19	0.07	-0.01	0.08	-0.11
Basin coverage	%	-0.02	0.91	-0.07	0.31	-0.01	-0.10	0.01	0.16	0.10
<b>Slope</b>										
Mean	degrees	0.03	0.25	-0.58	-0.49	-0.39	0.12	-0.33	0.21	-0.03
Sd	-	0.65	-0.22	-0.36	-0.29	0.41	-0.27	-0.13	0.11	0.01
Min	degrees	-0.48	0.56	0.16	-0.46	-0.11	0.32	0.29	-0.03	0.11
Max	degrees	0.94	0.05	-0.05	-0.13	0.23	0.01	-0.15	-0.11	0.08
Variety	-	0.97	0.07	0.01	-0.05	0.01	0.01	-0.09	-0.15	0.11
<b>Stream Order</b>										
1st order	%	0.52	0.18	0.62	0.36	-0.05	0.16	-0.23	-0.06	-0.29
2nd order	%	-0.23	0.45	0.38	0.38	0.51	0.14	-0.04	0.02	0.07
3rd order	%	0.14	-0.52	-0.10	-0.41	0.05	0.64	0.14	0.07	-0.13
4th order	%	-0.02	-0.37	-0.08	-0.04	-0.29	-0.85	0.03	-0.04	0.03
5th order	%	0.53	-0.20	-0.39	0.30	0.32	0.16	-0.20	0.32	0.22
6th order	%	0.39	-0.14	-0.19	0.53	-0.25	0.33	0.02	-0.13	0.05
7th order	%	0.93	0.18	0.20	-0.06	-0.04	-0.04	0.21	0.03	-0.08
8th order	%	0.89	0.17	0.20	-0.02	-0.05	-0.03	0.27	0.10	-0.17
<b>Elevation</b>										
Mean	m	0.49	0.31	-0.60	0.22	0.24	0.02	-0.11	-0.25	-0.31
Sd	-	0.91	-0.18	-0.24	-0.06	0.00	0.14	-0.11	0.09	-0.06
Min	m	-0.61	0.45	-0.11	0.16	0.38	-0.15	0.00	-0.36	-0.22
Max	m	0.92	0.08	-0.12	-0.11	0.14	-0.02	-0.24	-0.13	0.13
Variety	-	0.96	0.04	-0.03	-0.10	0.07	0.01	-0.11	-0.14	0.12
<b>Wetness index</b>										
Mean	-	0.38	-0.86	0.10	0.16	-0.10	-0.04	-0.03	0.21	0.03
Sd	-	0.63	-0.67	-0.35	0.11	-0.01	0.03	0.01	0.07	-0.01
Min	-	-0.85	-0.11	0.14	0.17	-0.16	-0.03	0.29	0.21	-0.13
Max	-	0.93	-0.21	-0.14	0.17	0.00	0.11	-0.02	-0.01	-0.06
Variety	-	0.95	0.17	0.17	-0.06	-0.02	-0.03	0.15	-0.01	-0.04
Proportion Positive	%	0.22	0.54	-0.28	-0.49	0.11	0.31	0.15	0.03	0.11
Proportion Negative	%	0.45	-0.36	0.70	0.28	-0.04	0.08	-0.10	0.13	-0.16
Eigenvalue	-	19.15	5.89	4.81	2.64	2.29	2.00	1.25	0.97	0.87

Table 4. Results of Principle Component Analysis for topographical and landcover variables in a 5km buffer around the sample site. 95% of the variance in the data was described with 6 components. Each component correlates differently to different variables, red highlights negative highlights, blue indicates positive correlations. Surface area is constant, at  $\pi 5^2=79\text{km}^2$ . PCA<sub>5km1</sub> represents sites surrounded by flat lowland areas and urban, agriculture and the flood plains of large rivers. PCA<sub>5km2</sub> represents sites surrounded by sloped highland areas and urban and agriculture, and small rivers. PCA<sub>5km3</sub> represents sites surrounded by sloped highland areas with savannah, and large swampy rivers. PCA<sub>5km4</sub> represents sites surrounded by flat lowland areas with savannah and small rivers. PCA<sub>5km5</sub> represents sites surrounded by flat highlands with urban and agriculture, and large rivers. PCA<sub>5km6</sub> represents sites surrounded by lowland hills, with small rivers and many small basins, in unforested environment.

Covariate	Units	PCA <sub>5km1</sub>	PCA <sub>5km2</sub>	PCA <sub>5km3</sub>	PCA <sub>5km4</sub>	PCA <sub>5km5</sub>	PCA <sub>5km6</sub>
<b>Land cover</b>							
Urban	%	0.72	0.40	-0.28	-0.42	0.10	0.08
Agriculture	%	0.84	0.20	-0.10	-0.08	0.30	0.06
Forest	%	-0.81	-0.53	-0.06	0.01	-0.15	-0.11
Roads	%	-0.03	-0.27	-0.35	0.02	-0.55	0.65
Savannah	%	-0.37	0.70	0.39	0.29	-0.14	0.20
Flood plain	%	0.96	0.20	-0.04	-0.07	0.02	-0.03
Swamp	%	0.02	-0.64	0.57	0.03	0.29	0.21
<b>Flow accumulation</b>							
Mean	m <sup>2</sup>	0.69	-0.26	0.57	-0.22	0.02	-0.14
Sd	-	0.70	-0.28	0.60	-0.20	0.03	-0.12
Max	m <sup>2</sup>	0.71	-0.29	0.60	-0.11	0.04	-0.10
Variety	-	0.73	0.02	0.25	0.58	0.06	0.15
<b>Basins</b>							
Mean basin depth	m	-0.95	-0.24	0.06	0.14	0.00	-0.11
Sd basin depth	-	0.94	0.10	0.11	-0.20	-0.12	0.15
Deepest basin	m	0.61	0.50	-0.31	0.06	-0.17	-0.27
Variety of basin depth	-	-0.53	-0.54	0.37	-0.21	0.09	0.32
Basin coverage	%	0.96	0.23	-0.05	0.02	0.03	0.07
<b>Slope</b>							
Mean	degrees	-0.89	0.06	0.19	-0.36	-0.06	0.11
Sd	-	-0.41	0.74	0.46	-0.13	-0.03	0.18
Max	degrees	-0.51	0.59	0.52	-0.06	-0.07	0.00
Variety	-	-0.66	0.48	0.42	-0.35	0.01	0.17
<b>Stream Order</b>							
1st order	%	0.93	0.32	-0.09	-0.04	0.03	0.01
2nd order	%	0.66	0.20	0.28	-0.12	0.32	0.07
3rd order	%	0.57	0.45	-0.51	-0.31	-0.24	0.03
4th order	%	0.40	-0.14	-0.20	0.19	0.54	0.51
5th order	%	-0.73	-0.19	-0.16	0.27	0.26	0.24
6th order	%	0.01	-0.48	0.60	0.54	-0.25	-0.08
7th order	%	0.16	0.78	0.17	0.50	-0.19	0.05
<b>Elevation</b>							
Mean	m	-0.83	0.36	0.02	0.00	0.30	-0.17
Sd	-	-0.21	0.79	0.44	-0.15	-0.09	0.07
Min	m	-0.56	0.57	-0.18	0.24	0.43	-0.19
Max	m	-0.66	0.67	0.24	-0.03	0.08	0.00
Variety	-	-0.65	0.66	0.33	-0.14	0.02	0.07
<b>Wetness index</b>							
Mean	-	0.98	0.09	0.09	0.12	-0.05	0.06
Sd	-	0.52	-0.18	0.77	0.24	-0.08	-0.05
Min	-	0.59	-0.71	-0.31	-0.13	0.03	-0.03
Max	-	0.77	-0.07	0.61	0.03	-0.04	-0.04
Variety	-	0.74	0.51	0.28	0.01	-0.03	0.04
Proportion Positive	%	-0.98	-0.07	0.07	-0.10	-0.02	-0.01
Proportion Negative	%	0.95	0.28	-0.05	0.00	0.04	0.05
Eigenvalue	-	18.79	8.02	5.49	2.17	1.51	1.29

### *Model fitting and evaluation*

We allow model selection to choose which of these principal components are most informative in the species distribution,  $Y_{\text{wet}}$  and  $Y_{\text{dry}}$ . The dry season general linear models (GLMs) and wet season GLMs were fitted separately with `glmulti` in the `glmulti` library in R. `Glmulti` finds the best set of GLMs among all possible combinations of explanatory variables; so for example all possible  $Y_{\text{dry}} \sim \text{PCA}_{5\text{km}}$  models were fitted, and each was evaluated with the Akaike information criterion corrected for small sample sizes (AICc). Low AICc scores indicate good performance and reduced overfitting (Warren and Seifert, 2011). The best set of these binomial GLMs (within 2 AICc scores of the best model) are selected, and the model within this range with the lowest sum of absolute residuals (best performance) is selected as the final model (Figure 24).

The response variable changed seasonally, resulting in two response variables,  $Y_{\text{dry}}$  and  $Y_{\text{wet}}$ . Along with the  $\text{PCA}_{5\text{km}}$  and  $\text{PCA}_{\text{ws}}$  inputs this resulted in four models;  $Y_{\text{dry}} \sim \text{PCA}_{5\text{km}}$  and  $Y_{\text{dry}} \sim \text{PCA}_{\text{ws}}$  in the dry season, and  $Y_{\text{wet}} \sim \text{PCA}_{5\text{km}}$  and  $Y_{\text{wet}} \sim \text{PCA}_{\text{ws}}$  in the wet season. This reduces our variables by retaining those that are important. Then, to compare the importance of  $\text{PCA}_{5\text{km}}$  (local) and  $\text{PCA}_{\text{ws}}$  (regional watershed) in the distribution of the response variable, *M. ulcerans* abundance, the components retained in these models were included in the final models,  $Y_{\text{dry}} \sim \text{PCA}_{5\text{km}} + \text{PCA}_{\text{ws}}$  in the dry season, and  $Y_{\text{wet}} \sim \text{PCA}_{5\text{km}} + \text{PCA}_{\text{ws}}$  in the wet season. In this way, by allowing `glmulti` to retain or drop these variables we can compare the importance of the watershed and local 5km area variables in the distribution of *M. ulcerans*.

Potential effects of multicollinearity were explored but were deemed minimal, as all pairwise Pearson correlation coefficient R values in the principal components were below 0.75 (Tables 5 and 6).

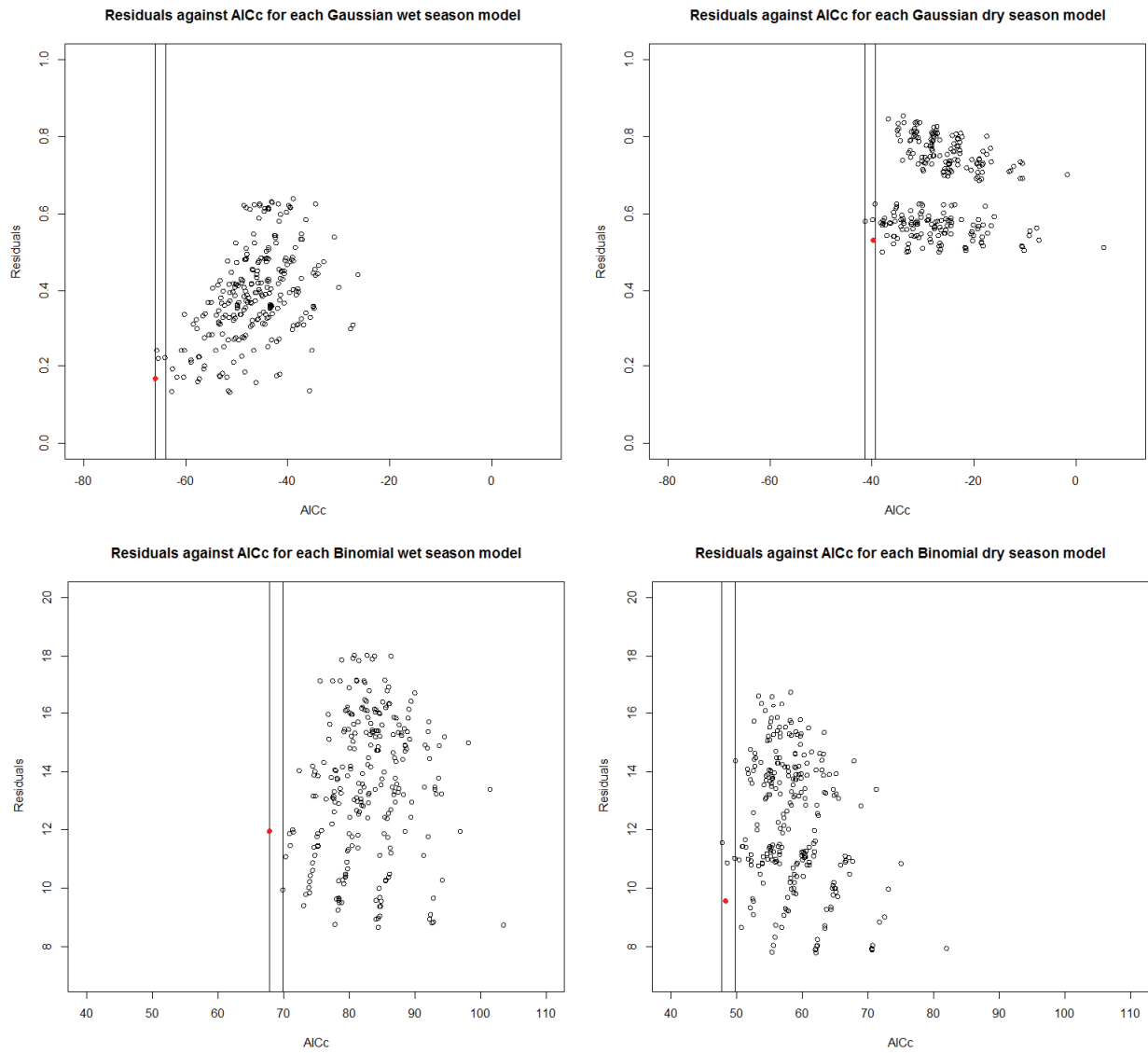


Figure 24. GLMulti output, for binomial and Gaussian models. Sum of absolute model residuals are plotted against AICc. Within the region of 2 AICc scores of the best model (vertical lines) we select the model with the lowest residuals (highlighted in red).

Table 5. Pearson product R correlation coefficients in the wet season model. Stepwise selection selected 3 components, none of which were correlated.

	PCAws4	PCAws9	PCA5km2
PCAws4	1	0	0.13
PCAws9	0	1	-0.16
PCA5km2	0.13	-0.16	1



Table 6. Pearson product R correlation coefficients in the dry season model. Stepwise selection selected 6 components, none of which were correlated.

	PCAws1	PCAws5	PCAws6	PCAws8	PCA5km2	PCA5km4
PCAws1	1	0	0	0	-0.15	-0.03
PCAws5	0	1	0	0	0.71	0.25
PCAws6	0	0	1	0	0.11	0.08
PCAws8	0	0	0	1	0.11	-0.48
PCA5km2	-0.15	0.71	0.11	0.11	1	
PCA5km4	-0.03	0.25	0.08	-0.48	0	1

In the initial screen of variables,  $Y_{dry} \sim PCA_{5km}$  and  $Y_{dry} \sim PCA_{ws}$  retained  $PCA_{ws}4$ , “small watersheds that drain swamp and forest at flat intermediate elevations”,  $PCA_{ws}9$ , “small watersheds that drain wet swamps in areas that reach from low to high elevations” and  $PCA_{5km}2$ , “sites surrounded by sloped highland areas with urban, agriculture and small rivers”. These were included in the model of interest,  $Y_{dry} \sim PCA_{5km} + PCA_{ws}$ .

For the wet season  $Y_{wet} \sim PCA_{5km}$  and  $Y_{wet} \sim PCA_{ws}$  retained  $PCA_{ws}1$ , “large watersheds that drain flood plains”,  $PCA_{ws}5$ , “small watersheds that drain highland urban and savannah”,  $PCA_{ws}6$ , “small watersheds that drain highland urban and forest”,  $PCA_{ws}8$ , “small watersheds that drain urban and agricultural environments in hilly lowlands”,  $PCA_{5km}2$ , “sites surrounded by sloped highland areas with urban, agriculture and small rivers” and  $PCA_{5km}4$ , “sites surrounded by flat lowland areas with savannah and small rivers”, which were included in  $Y_{wet} \sim PCA_{5km} + PCA_{ws}$ .

*Predicting the spatial distribution of suitable habitat for M. ulcerans in the model training region, Akonolinga*

We interpolate the Akonolinga model within the region of Akonolinga to predict the distribution of suitable habitat, the reservoir, of *M. ulcerans*. To achieve this, points where streams (defined using STRM) flow under or across roads (defined using satellite images) were selected. These were termed ‘pour points’ in this article. Selection of the point where streams cross roads was based on the hypothesis that these environments, where contact between humans and the aquatic environment will be high, may be important in infection. This does not mean that infection does not occur in other locations, nor do we speculate on the importance of relative routes of transmission. This will not characterise all the environmental reservoir of the bacterium, but will describe an important part of it. The topography and land cover of the watershed and 5km buffer of these pour points was

characterised, transformed into  $PCA_{5km}$  and  $PCA_{ws}$  format, and the GLM was predicted. As a summary to describe this distribution, we use Morans Index of spatial autocorrelation, which describes the extent to which the distribution is random, and is here used to describe the distribution of suitable sites. This is implemented using the tool Spatial Autocorrelation Global Moran's I in ArcMap10.1.

*Predicting the spatial distribution of suitable habitat for M. ulcerans in a new region, French Guiana*

We extrapolate the Akonolinga wet season model to French Guiana, to understand how the suitable habitat in one region is similar to that in another. For comparability, the wet season model, constructed in Cameroon, was used to predict the positive sites among the 18 sampled sites in French Guiana. Values of  $PCA_{5km}$  and  $PCA_{ws}$  in French Guiana were generated using the ind.sup option in the PCA function. The Akonolinga wet season model was then predicted into French Guiana using the land cover data provided by the French *Ministère de l'Écologie, du Développement Durable et de l'Énergie* (CORINE Land Cover, 2006), and topography derived from SRTM.

As discussed above, the choice of error structure is important in the performance of a GLM. We aim to describe the distribution of the bacterium, so preference is given to the model with the lowest residual values in the model, which in this case is Gaussian rather than Binomial error structure. Residuals were much lower in a Gaussian model, as shown in Supplementary Figures 2 and 3 (see the observed response versus predicted response for Gaussian and Binomial models and QQ plots for the Gaussian and Binomial models, respectively). This difference is an order of magnitude. This was a practical decision – using Gaussian models in this case was based entirely on the desire to clearly predict where this pathogenic bacterium is more likely to occur, in such a case errors of residuals have a greater cost.

The wet and dry season watershed Gaussian models were predicted on the pour point data using the predict.glm function in R. The model predictions of habitat suitability at these pour points were then interpolated using Inverse Distance Weighting in the IDW tool of ArcMap 10 (ESRI).

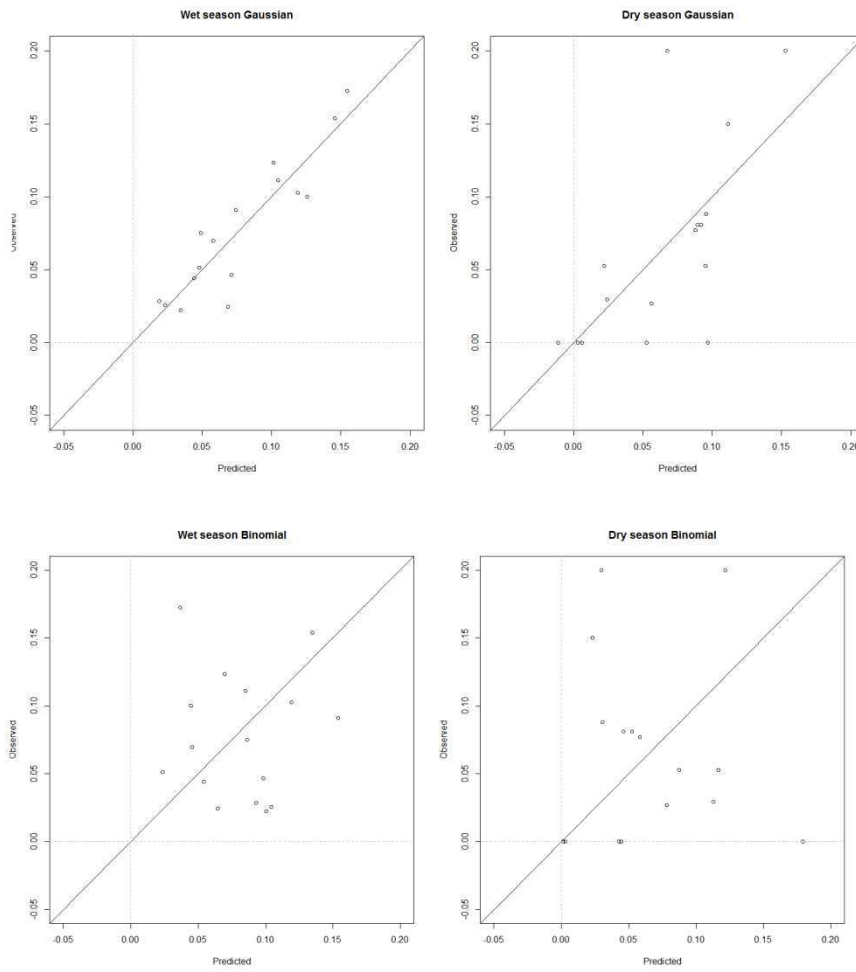


Figure 25. Observed against predicted values for each model. Note that Gaussian models have a much better fit.

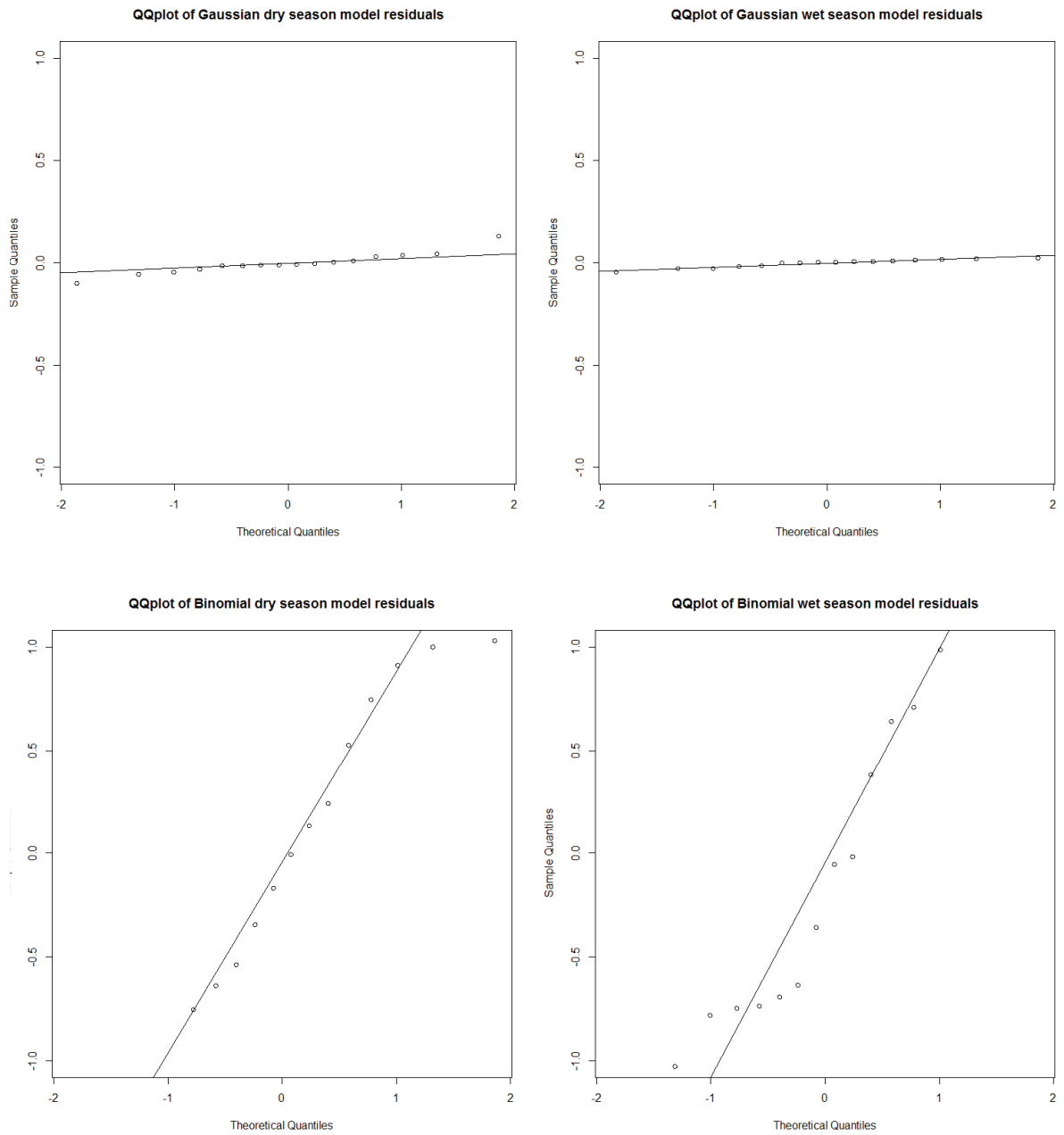


Figure 26. Quantile-Quantile plots of normality. The Gaussian and Binomial are both similarly normally distributed, though the Binomial displays a larger variance of residuals.

## Results

### *Relative importance of local and regional effects on the distribution of M. ulcerans in wet season*

The final fitted wet season Binomial logit GLM, after stepwise AICc selection, was

$$Y_{\text{wet}} \sim 1 + \text{PCA}_{\text{ws}9} + \text{PCA}_{5\text{km}2}$$

The final GLM suggested that both local and regional effects are substantially correlated to *M. ulcerans* distribution. Regional effects were represented by  $\text{PCA}_{\text{ws}9}$ , “small watersheds that drain wet swamps in areas that reach from low to high elevations”, and was negatively correlated to *M. ulcerans* abundance (correlation coefficient -0.37,  $p=0.007$ ). This means we expect less *M. ulcerans* in small watersheds that drain swamps near highlands. The second part of the above equation corresponds to local effects;  $\text{PCA}_{5\text{km}2}$  represents “sites surrounded by sloped highland areas with urban, agriculture and small rivers”. This was also negatively correlated to *M. ulcerans* abundance (correlation coefficient -0.16,  $p=0.00214$ ), so we expect less *M. ulcerans* when the area around the sample site is highland areas with urban and agricultural areas.

The spatial distribution of *M. ulcerans* suitable habitat in the wet season predicted at the pour points was non-random, based on Moran’s I spatial autocorrelation (Moran's Index: 0.21, z-score: 9.1,  $p<0.00001$ ), positive sites tend to cluster together (Figure 27).

### *Relative importance of local and regional effects on the distribution of M. ulcerans in dry season*

The final fitted dry season binomial logit GLM, after stepwise AICc selection, is

$$Y_{\text{dry}} \sim 1 + \text{PCA}_{\text{ws}1} + \text{PCA}_{5\text{km}2} + \text{PCA}_{5\text{km}4}$$

The final models on the dry season found that both regional and local effects were substantially correlated to presence of *M. ulcerans*. Regional effects were represented by  $\text{PCA}_{\text{ws}1}$ , “large watersheds that drain flood plains”, which was marginally negatively correlated to *M. ulcerans* abundance (correlation coefficient -0.26,  $p=0.05210$ ).  $\text{PCA}_{5\text{km}2}$ , “sites surrounded by areas with urban, agriculture and small rivers” was positively correlated to *M. ulcerans* abundance (correlation coefficient 0.09,  $p=0.18709$ ) though the  $p$  value suggests this is not significant, and finally  $\text{PCA}_{5\text{km}4}$ , “sites surrounded by areas with savannah and small rivers”, was positively correlated to *M. ulcerans* abundance, (correlation coefficient 0.38,  $p=0.007$ ).

The spatial distribution of *M. ulcerans* suitable habitat in the dry season predicted at the pour points is non-random, based on Moran's I spatial autocorrelation (Moran's Index: 0.33, z-score: 14.32,  $p < 0.00001$ ) positive sites tend to cluster together (Figure 27).

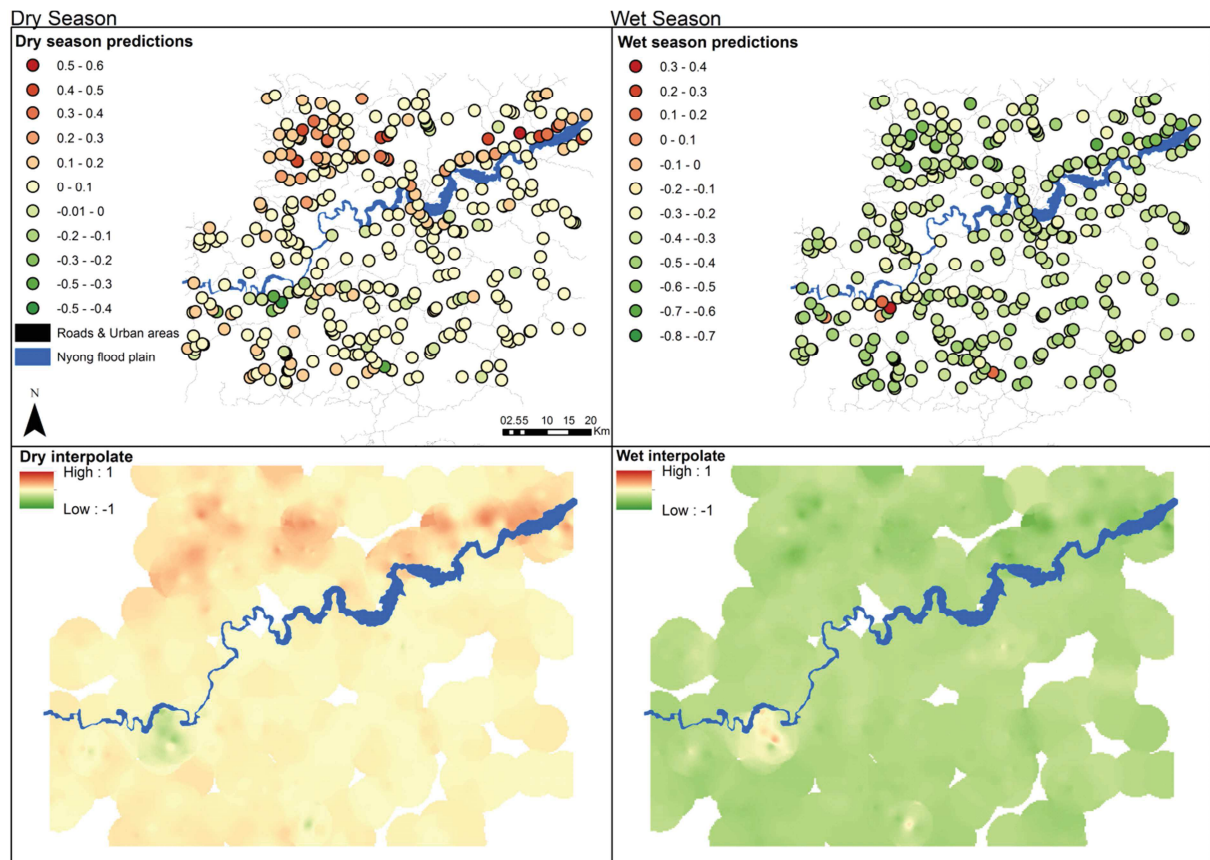


Figure 27. Spatial distribution of habitat suitable for *M. ulcerans* in Akonolinga, Cameroon. Units of habitat suitability are the proportion of qPCR pools predicted to be positive, based on the field work of Garchitorea *et al.* (2014). Negative values are a result of the normal distribution of the residuals (Supplementary Figures 4 and 5). The Gaussian wet and dry season models, based on the original 16 sites, are predicted into each of the pour points (where a stream crosses a road) in the region (top row), resulting in the predicted habitat suitability at each point. The pour points are interpolated (bottom row) using IDW fixed distance 0.05 decimal degrees interpolation (ArcMap10.1) resulting in the first map of spatial distribution of *M. ulcerans* encounter risk.

#### Model performance when interpolated in Akonolinga

Spatial autocorrelation of model residuals can be an issue in GLMs, but this was explored, and it was not the case here. Model residuals were not significantly spatially autocorrelated in the wet season (Moran's Index: -0.285386, z-score: -1.045844,  $p = 0.295633$ ) nor in the dry season (Moran's Index: 0.071225, z-score: 0.655435,  $p = 0.512187$ ).

The AICc of the final dry season Binomial model was 49.6, the absolute sum of the residuals was 11.03. The AICc of the final wet season Binomial model was 67.8, the absolute sum of the residuals was 11.95.

We note that Gaussian models had significantly better performance. The AICc of the final dry season Gaussian model was -39.8, the absolute sum of the residuals was 0.53. The AICc of the final wet season Gaussian model was -65.5, the absolute sum of the residuals was 0.24. Model performance is presented in Figure 25, model residuals were normally distributed (Figure 26).

*Model performance when extrapolated in French Guiana*

The Akonolinga wet season model was predicted into 18 sample sites in French Guiana (Figure 28, 2<sup>nd</sup> row). The model predicted sites to be positive or negative, and the results of qPCR corroborated these predictions (Figure 28). Performance of the Binomial model was notably poor, all sites were predicted negative. In contrast, performance of the Gaussian model was better, but accuracy was still poor at 0.39 (Table 7). Sensitivity and negative predictive values are high, indicating that the predictions of presence of the bacterium are likely to be true, specificity and positive predictive values are low; indicating predictions of absence of the bacterium are likely to be incorrect. This is a result of a bias towards Type II errors (false negatives) in the Gaussian model. Overall, the model predicts *M. ulcerans* in Akonolinga, but is sensitive to extrapolation. Extrapolation tends to result in false negative predictions of presence.

Table 7. Contingency table describing model performance of niche models constructed in Cameroon and predicted into French Guiana. The rows 'Prediction' are model predictions, 'Test' are the results from qPCR of the sites in French Guiana. Values in blue are true positives and true negatives; values in red are false positives and false negatives.

		Test				
		positive	negative			
Prediction	positive	3	2	Sensitivity	0.60	
	negative	9	4	Specificity	0.31	
	PPV	0.25	NPV	0.66	Accuracy	0.39

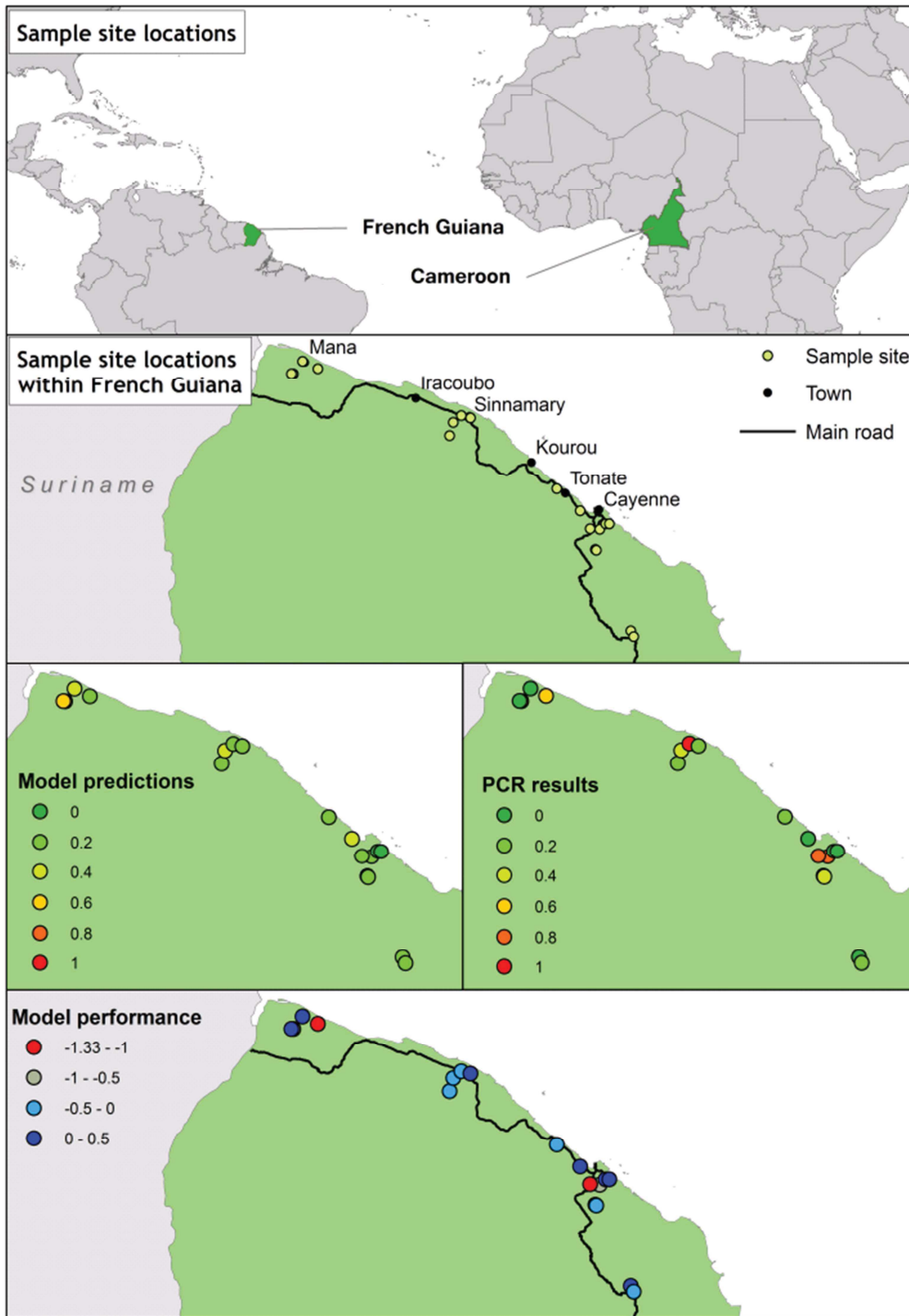


Figure 28. Model validation in French Guiana. Sample sites were as in Morris *et al.* (2014). A wet season Gaussian niche model based on data collected in Cameroon was predicted into French Guiana (3<sup>rd</sup> row, left hand side). The model under-predicted, *M. ulcerans* was present in more sites than expected (bottom row, model residuals). A similar Binomial model predicted all sites to be negative.



## Discussion

Here, we have demonstrated that in addition to local variables around the sample site, the distribution of *M. ulcerans* correlates to regional variables, i.e. the topography and land cover of the watershed of the sample site. This spatial distribution of suitable habitat was described, allowing the production of environmental hazard maps for the distribution of the pathogen. *M. ulcerans* presence in the wet season correlates with lowland areas surrounded by few agricultural or urban areas, particularly if the sample site has a large watershed. We expect more *M. ulcerans* in the dry season in sites surrounded by urban and agricultural areas, with many small streams, particularly if the sample site has a small watershed.

Many of the findings are in accord with what little we already understand about this bacterium. *M. ulcerans* has been previously associated with flat wetland areas (Johnson *et al.* 2005, Wagner *et al.* 2008). A similar association with Buruli ulcer was reported by Wagner *et al.* (2008 B), who found that high standard deviation of the wetness index was a risk factor for Buruli ulcer. These three variables are normally strongly correlated to each other and ecologically similar entities. In this study these are negatively correlated to PCA<sub>ws9</sub>, here termed “small watersheds that drain wet swamps in areas that reach from low to high elevations” which negatively correlated to *M. ulcerans* abundance: these studies appear to be describing the same ecological entity, but with different variables.

Our study was limited in certain regards, as we focused it on the prevalence of *M. ulcerans* in the biotic community, and on how topography and land cover in the region could influence that prevalence. We do not consider abiotic conditions testing positive for *M. ulcerans*. Potentially the abiotic distribution may respond differently to these variables, future work will aim to explore this. However, given that *M. ulcerans* is commonly detected in the biotic environment and appears to be at lower prevalence in the abiotic environment, we believe our results are still applicable to an understanding of *M. ulcerans* distribution. We had a relatively low positivity rate (Table 1). A potential limitation is that low positivity can bias a model towards false negatives, while this is possible we are unable to test this further with our current data.

The Akonolinga wet season model was extrapolated into French Guiana, where sampling was in the wet season. Despite good performance in Akonolinga, the model performed poorly in French Guiana, under-predicting the bacterium’s distribution (Figure 6). There are a number of points to be drawn from this. First, there were differences in sampling effort between the two sites, as the Akonolinga sampling regime consisted of 12 time points in the year, while the French Guiana regime consisted of 2 time points. This would be consistent with the idea that the bacterium is transiently

present in different regions, and under-prediction would be expected in this case. Secondly, a potential complication results from differences in the ability of the SRTM dataset to delineate watersheds due to dense rainforest canopies in French Guiana (Roux *et al.* 2010). The shape of a watershed is sensitive to the quality of the elevation data used, errors in the digital elevation model, or man-made drainage structures, can have effects not captured by this model. Finally, we cannot rule out that the differences are a result of differences in *M. ulcerans*. We used qPCR to detect *M. ulcerans*, however the species is known to have multiple ecovars (Vandelannoote *et al.* 2013, Tobias *et al.* 2013) and subspecies, distributed differently throughout the globe. If it is the case that we are predicting the ecological niche of one Akonolinga *M. ulcerans* species into French Guiana, and testing it against a separate French Guiana species, one would expect the model to under-predict if the French Guiana subspecies occupies a larger ecological niche.

Regardless of error structure, selection of both types of models (Gaussian and Binomial) retained watersheds as important variables. These findings will impact future research on Buruli ulcer and *M. ulcerans*; future sampling regimes would benefit by consideration of the local hydrology before beginning sampling, and selecting sample sites along these lines. We also postulate the importance of watersheds as a barrier to dispersal for the bacterium. In a recent key study Vandelannoote *et al.* (2013) found a strong relationship between *M. ulcerans* population structure and the greater West African hydrological watersheds, with populations being bound to watersheds. These are the drainage areas of large rivers such as the Nyong, Mbam and Ouémé rivers, a much larger scale than our study. However, given our results herein, it seems the bacteria may drift downstream. This is inferred by the difference in the effect of watershed size from dry to wet seasons.

This is consistent with the idea of a 'flushing' effect of rainfall in the wet season, carrying bacteria downstream Morris *et al.* (2014B), which will influence their genetic population structure. This has notable consequences for the epidemiology of Buruli ulcer. If the watersheds are barriers to movement for the bacteria it implies that *M. ulcerans* may be common in the environment, but in certain areas hydrological conditions facilitate concentration of the bacterium, as is the case with anthrax (Dragon and Rennie 1995).

## Conclusion

The distribution of environmental pathogens needs to be understood to facilitate control. Commonly, local effects in the microhabitats are considered to describe the ecological niche of a pathogen. However our study demonstrates that regional effects are important factors to be considered. Future research on the *M. ulcerans* would benefit by considering the watershed of potential sample sites, particularly as such data is often quite simple to acquire. The shape, size, and land cover of the watershed correlates with changes in the distribution of *M. ulcerans*, and useful information is lost if watersheds are ignored. The distribution of swamp in a watershed was found to be an important factor in the suitability of the site for *M. ulcerans*; though a sample point in the field may be at a location normally considered unsuitable for the bacteria (e.g. a small swift lentic stream), the area upstream may contain an abundance of lotic swamps and be quite suitable for the bacterium, which may be 'washed out' downstream towards the sample site. This is an example of the useful information we gain by placing pathogens in an environmental context, rather than regarding them solely in an epidemiological sense.

## References

- Ari, T. B., *et al.* (2011) Plague and climate: scales matter. *PLoS pathogens* 7.9: e1002160.
- Benbow, E., *et al.* (2008) Aquatic invertebrates as unlikely vectors of Buruli ulcer disease. *Emerging infectious diseases* 14.8: 1247.
- Benbow, E., *et al.* (2013) Aquatic Macroinvertebrate Assemblages of Ghana, West Africa: Understanding the Ecology of a Neglected Tropical Disease. *Ecohealth* 1-16.
- Beven, K. and Kirkby, M. (1979) A physically based, variable contributing area model of basin hydrology *Hydrological Sciences Bulletin*, 24, 43–69, 1979
- Carson, C., *et al.* (2014) Potential Wildlife Sentinels for Monitoring the Endemic Spread of Human Buruli Ulcer in South-East Australia. *PLoS neglected tropical diseases* 8.1: e2668.
- Dragon, D., and Rennie R. (1995) The ecology of anthrax spores: tough but not invincible. *The Canadian Veterinary Journal* 36.5: 295.
- eCognition Trimble Navigation Ltd, Arnulfstrasse, Munich Germany
- Eddyani, M., *et al.* (2004) Potential role for fish in transmission of *Mycobacterium ulcerans* disease (Buruli ulcer): an environmental study. *Applied and environmental microbiology* 70.9: 5679-5681.
- ESRI 2011. ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute.
- French Ministère de l'Écologie, du Développement Durable et de l'Énergie CORINE Land Cover 2006
- Fyfe, J., *et al.* (2007) Development and application of two multiplex real-time PCR assays for the detection of *Mycobacterium ulcerans* in clinical and environmental samples. *Applied and environmental microbiology* 73.15: 4733-4740.
- Fyfe, J., *et al.* (2010) A major role for mammals in the ecology of *Mycobacterium ulcerans*. *PLoS neglected tropical diseases* 4.8: e791.
- Garchitorena, A., *et al.* (2014) *Mycobacterium ulcerans* Ecological Dynamics and Its Association with Freshwater Ecosystems and Aquatic Communities: Results from a 12-Month Environmental Survey in Cameroon. *PLoS neglected tropical diseases* 8.5: e2879.
- Hay, S., *et al.* (2013) Global mapping of infectious disease. *Philosophical Transactions of the Royal Society B: Biological Sciences* 368.1614: 20120250.
- Hutchinson, G.E. (1957) Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology* 22 (2): 415–427.
- Jarvis, A., *et al.* (2008) "Hole-filled SRTM for the globe Version 4", available from the CGIAR-CSI SRTM 90m Database (<http://srtm.csi.cgiar.org>)
- Johnson PDR, *et al.* (2005) Buruli Ulcer (*M. ulcerans* Infection): New Insights, New Hope for Disease Control. *PLoS Med* 2(5): e173. doi:10.1371/journal.pmed.0020173

- Johnson, PDR, *et al.* (2007) Mycobacterium ulcerans in mosquitoes captured during outbreak of Buruli ulcer, southeastern Australia. *Emerg Infect Dis* 13.11: 1653-1660.
- Kotlowski, R., *et al.* (2004) One-tube cell lysis and DNA extraction procedure for PCR-based detection of Mycobacterium ulcerans in aquatic insects, molluscs and fish. *Journal of medical microbiology* 53.9: 927-933.
- Marsollier, L., *et al.* (2002) Aquatic insects as a vector for Mycobacterium ulcerans. *Applied and environmental microbiology* 68.9: 4623-4628.
- Marsollier, L., *et al.* (2004 A) Aquatic plants stimulate the growth of and biofilm formation by Mycobacterium ulcerans in axenic culture and harbour these bacteria in the environment. *Applied and environmental microbiology* 70.2: 1097-1103.
- Marsollier, L., *et al.* (2004 B) Aquatic snails, passive hosts of Mycobacterium ulcerans. *Applied and environmental microbiology* 70.10: 6296-6298.
- Marion, E., *et al.* (2010) Seasonal and regional dynamics of *M. ulcerans* transmission in environmental context: deciphering the role of water bugs as hosts and vectors. *PLoS neglected tropical diseases* 4.7: e731.
- Merritt, R., *et al.* (2010) Ecology and transmission of Buruli ulcer disease: a systematic review. *PLoS neglected tropical diseases* 4.12: e911.
- McIntosh, M., *et al.* (2014) Associations between Mycobacterium ulcerans and aquatic plant communities of West Africa: implications for Buruli ulcer disease. *EcoHealth*: 1-13.
- Mosi, L., *et al.* (2008) Persistent association of Mycobacterium ulcerans with West African predaceous insects of the family Belostomatidae. *Applied and environmental microbiology* 74.22: 7036-7042.
- Morris, A., *et al.* (2014 A) First Detection of Mycobacterium ulcerans DNA in Environmental Samples from South America. *PLoS neglected tropical diseases* 8.1: e2660.
- Morris, A., *et al.* (2014 B) Complex temporal climate signals drive the emergence of human water-borne disease. *Emerging Microbes and Infections* 3, e56
- Morris, L., *et al.* (2013) Informing surveillance for the lowland plague focus in Azerbaijan using a historic dataset. *Applied Geography* 45: 269-279.
- Palomino, J., *et al.* (1998) Effect of oxygen on growth of Mycobacterium ulcerans in the BACTEC system. *Journal of clinical microbiology* 36.11: 3420-3422.
- Portaels, F., *et al.* (2008) First cultivation and characterization of Mycobacterium ulcerans from the environment. *PLoS neglected tropical diseases* 2.3: e178.
- Portaels, F., *et al.* (1999) Insects in the transmission of Mycobacterium ulcerans infection. *The Lancet* 353.9157: 986.

Portaels, F., *et al.* (2001) *Mycobacterium ulcerans* in wild animals. *Revue scientifique et technique (International Office of Epizootics)* 20.1: 252.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>

Roche, B., *et al.* (2013) Identifying the Achilles heel of multi-host pathogens: the concept of keystone 'host' species illustrated by *Mycobacterium ulcerans* transmission. *Environmental Research Letters* 8.4: 045009.

Soberon, J. (2005) "Interpretation of models of fundamental ecological niches and species' distributional areas."

Stinear, T., *et al.* (2007) Reductive evolution and niche adaptation inferred from the genome of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer. *Genome research* 17.2: 192-200.

Stinear, T., *et al.* (2000) Identification of *Mycobacterium ulcerans* in the environment from regions in Southeast Australia in which it is endemic with sequence capture-PCR. *Applied and environmental microbiology* 66.8: 3206-3213.

Strahler, A. N. (1957), "Quantitative analysis of watershed geomorphology", *Transactions of the American Geophysical Union* 38 (6): 913–920.

Tobias, N., *et al.* (2013) "Complete genome sequence of the frog pathogen *Mycobacterium ulcerans* ecovar Liflandii." *Journal of bacteriology* 195.3: 556-564.

Tran, A., *et al.* (2008) Using remote sensing to map larval and adult populations of *Anopheles hyrcanus* (Diptera: Culicidae) a potential malaria vector in Southern France. *International Journal of Health Geographics*. 7:9.

Trott, K., *et al.* (2004) Characterization of a *Mycobacterium ulcerans*-like infection in a colony of African tropical clawed frogs (*Xenopus tropicalis*). *Comparative medicine* 54.3: 309-317.

Vandelannoote, K, *et al.* (2014) Insertion Sequence Element Single Nucleotide Polymorphism Typing Provides Insights into the Population Structure and Evolution of *Mycobacterium ulcerans* across Africa. *Applied and environmental microbiology* 80.3: 1197-1209.

van Ravensway, J, *et al.* (2012) Climate and landscape factors associated with Buruli ulcer incidence in Victoria, Australia. *PloS one* 7.12: e51074.

Walsh, D., *et al.* (2008) Buruli ulcer (*Mycobacterium ulcerans* infection). *Transactions of the Royal Society of Tropical Medicine and Hygiene* 102.10: 969-978.

Warren, D., and Seifert S. (2011) Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecological Applications* 21.2: 335-342.

Wagner, T., *et al.* (2008 A) A landscape-based model for predicting *Mycobacterium ulcerans* infection (Buruli ulcer disease) presence in Benin, West Africa. *EcoHealth* 5.1: 69-79.

Wagner, T., *et al.* (2008 B) Buruli ulcer disease prevalence in Benin, West Africa: associations with land use/cover and the identification of disease clusters *International journal of health geographics* 7: 25.

WHO (2008) Buruli ulcer: progress report, 2004–2008. Geneva, Switzerland.

Williamson, H., *et al.* (2008) "Distribution of *Mycobacterium ulcerans* in Buruli ulcer endemic and non-endemic aquatic sites in Ghana." *PLoS neglected tropical diseases* 2.3: e205.

Williamson, H., *et al.* (2012) "Detection of *Mycobacterium ulcerans* in the environment predicts prevalence of Buruli ulcer in Benin." *PLoS neglected tropical diseases* 6.1: e1506.

Williamson, H., *et al.* (2014)"*Mycobacterium ulcerans* Fails to Infect through Skin Abrasions in a Guinea Pig Infection Model: Implications for Transmission." *PLoS neglected tropical diseases* 8.4: e2770.

**Chapter 4: Ecological niche modelling of Hemipteran insects in  
Cameroon; the paradox of a vector-borne transmission for  
*Mycobacterium ulcerans*, the causative agent of Buruli ulcer**



## Foreword.

The following chapter is currently under review as a paper to the International Journal of Health Geographics, titled “Ecological niche modelling of Hemipteran insects in Cameroon; the paradox of a vector-borne transmission for *Mycobacterium ulcerans*, the causative agent of Buruli ulcer.”, by Kevin Carolan, Solange Meyin À Ebong, Andres Garchitorena, Jordi Landier, Daniel Sanhueza, Gaëtan Texier, Laurent Marsollier, Philippe Le Gall, Jean-François Guégan and Danny Lo Seen.

This has involved extensive collaboration with Solange Meyin À Ebong, who is supported by the Centre Pasteur du Cameroun and the MIVEGEC lab of IRD, Jordi Landier, who is supported by a PhD studentship from Institut Pasteur, Unité d'Epidemiologie de Maladies Emergentes, and Daniel Sanhueza, who is supported by a PhD studentship from *Becas* Agencia de Cooperación Internacional de Chile (AGCI) and the BIOHOPSYS programme from the LabEx-CEBA.

This work has explored the possibility that these Hemipteran insects (Figure 13) are likely to act as vectors of Buruli ulcer. We expect that, under most circumstances, the abundance of a vector should be correlated to the prevalence of the disease. This assumption is discussed in the paper, and an initial criticism of this work will be that correlation does not imply causation. Of course this is true, and we do not herein seek to conclusively end the discussion on the role of these insects as vectors, but rather to contribute to the debate.

Herein, the extent of the study is limited using a technique called multivariate environmental similarity surfaces (MESS). The idea of using this to limit the extent of the study was developed following the work in chapter 2. The concept is explained in Figure 29, below; a MESS map describes the similarity of a given set of sample sites to the rest of the study area, for a given set of environmental variables. The study is then limited to the region ecologically interpolated by the sample sites.

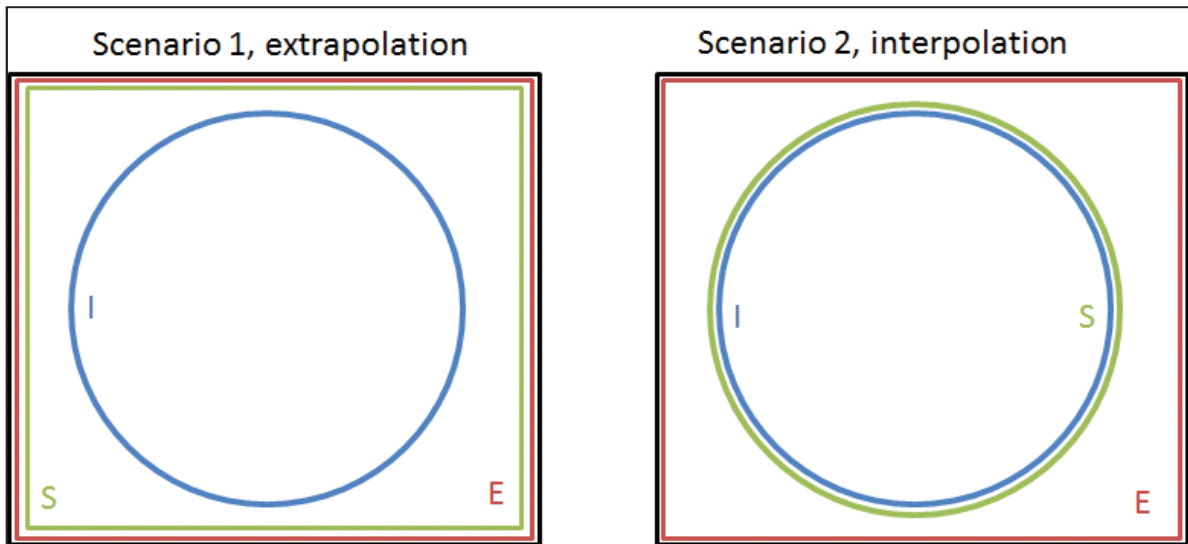


Figure 29. BAM diagram explaining the method used to limit the area of study. **I** is the area interpolated by the sample sites, **E** is the area extrapolated beyond the sample sites, and **S** is the area considered in the modelling method. If the area of the study is limited such that we extrapolate, that is that  $S=E$ , we make predictions that are unjustified based on our information. Instead, we limit the study such that  $S=I$ , using MESS maps to identify the point where we exceed **I**. This limits the predictive power of our model to a certain region, with more reliable predictions within that region.

## **Abstract**

### *Background*

The mode of transmission of the emerging neglected disease Buruli ulcer is unknown. Several potential transmission pathways have been proposed, such as amoebae, or transmission through food webs. Several lines of evidence have suggested that biting aquatic insects, Naucoridae and Belostomatidae, may act as vectors, however this proposal remains controversial.

### *Materials and Methods*

Herein, based on sampling in Cameroon, we construct an ecological niche model of these insects to describe their spatial distribution. We predict their distribution across West Africa, describe important environmental drivers of their abundance, and examine the correlation between their abundance and Buruli ulcer prevalence in the context of the Bradford-Hill guidelines.

### *Results*

We find a significant positive correlation between the abundance of the insects and the prevalence of Buruli ulcer. This correlation changes in space and time, it is significant in one Camerounese study region in (Akonolinga) and not the other (Bankim). We discuss notable environmental differences between these regions.

### *Conclusion*

We interpret the presence of, and change in, this correlation as evidence (though not proof) that these insects may be locally important vectors. This is consistent with the idea of Buruli ulcer as a disease transmitted by multiple modes of infection, the importance of any one pathway changing from region to region, depending on the local environmental conditions.

Key-words: Ecological niche modelling, Naucoridae, Belostomatidae, spatial distribution, habitat suitability, Buruli ulcer, *Mycobacterium ulcerans*, vector-borne transmission, environmentally-acquired disease.

## Background

The Buruli ulcer is an emerging neglected tropical disease affecting more than 5000 people per year in West and Central Africa, French Guiana, Latin America and Australia (Van der Werf 1999, WHO). The disease burden is highest in Africa, where it predominantly affects children under the age of 15, and due to damage to the skin, muscle and bone, can result in severe scarring and crippling deformities if left untreated. The disease is caused by the environmental pathogen *Mycobacterium ulcerans*.

The mode of transmission of *M. ulcerans*, the method by which it infects humans, is unknown. Many routes of transmission have been proposed, such as transmission by aerosol (Veitch *et al.* 1997), vector transmission by amoebae (Amissah *et al.* 2014) or through aquatic networks (Roche *et al.* 2013). During a study of the association between *M. ulcerans* and aquatic plants in Ghana and Benin aquatic insects were accidentally collected during the sampling procedure, and unexpectedly found to test positive for *M. ulcerans* (Van der Werf *et al.* 1999). The authors proposed that, given that these insects occasionally bite humans, they may be implicated in transmission of *M. ulcerans*. Aquatic insects have been further implicated after a series of laboratory experiments demonstrated the competency of Naucoridae to act as vectors. Naucoridae are able to acquire *M. ulcerans* from their diet, and then transmit the pathogen to mice resulting in Buruli-like symptoms (Marsollier *et al.* 2002, Marsollier *et al.* 2005, Marsollier *et al.* 2007A, Marsollier *et al.* 2007B). Buruli ulcer is commonly associated with lowland, stagnant water (Wagner *et al.* 2008) and human behaviours associated with water bodies appear to be risk factors for Buruli ulcer infection, which would lend support to the idea of infection occurring in an aquatic context.

However, the role of these insects has been disputed for several reasons. In a two year study of Buruli ulcer endemic and non-endemic sites in Ghana, Benbow *et al.* (2008) found no evidence for a role in transmission. The population of these insects, and the prevalence of *M. ulcerans* infection in them, was not significantly different between Buruli ulcer endemic and Buruli ulcer non-endemic sites. The authors argued that, if these insects are vectors, we would expect them to have a higher abundance in Buruli ulcer endemic sites, that rates of *M. ulcerans* infection of the insects should be higher in Buruli ulcer endemic sites, and that the rate of infection of these insects should be higher than other species of invertebrates.

These expectations are based on the Bradford-Hill guidelines for associating insect vectors with human vector-borne disease (Hill 1965, Merritt *et al.* 2010). These guidelines provide a general

framework to explore the association between vectors and disease, based on the consistency, specificity, plausibility and coherence of the proposed mode of transmission.

Consistency refers to the expectation that the rate of infection of the proposed vectors, which should be consistently strongly positively associated, in time and space, with prevalence of human cases. This also implies human cases should not occur in absence of the proposed vector. The proposed vector should have a demonstrated capacity to physically transmit the pathogen, which has been demonstrated for *M. ulcerans* in the lab (Marsollier *et al.* 2002). The interaction between the proposed vector and human infection must be specific and alternative explanations of human infection should be ruled out (though see Roche *et al.* 2013 for alternative explanations). That is, human infection must be demonstrated to not have been the result of other potential modes of transmission. We note that this criterion must be applied with care in cases of multi-vectorial transmission.

Additionally, the proposed vector must plausibly be able to be a vector of the pathogen. This criterion is often controversial as it is highly dependent upon the experience of the researcher, and their opinion about what is, and is not, plausible as opposed to merely possible (Schünemann *et al.* 2011). Most authors in Buruli ulcer research would agree with some basic facts; the waterbugs are infected in the environment (Marion *et al.* 2010), they bite humans occasionally, and are able to transmit the bacteria to mice in the lab (Marsollier *et al.* 2002). However, waterbugs are not known to be vectors of other pathogens, and related Mycobacteria (*Mycobacterium tuberculosis*, *M. leprae*, *M. marnium*) are not known to be vector-borne diseases (Benbow *et al.* 2008, Merritt *et al.* 2010). The plausibility of this proposed route of transmission is still debated. The final criterion, coherence, is based on what we already know about the pathogen, the vector and the host. Does the proposed method of transmission fit well with our current understanding of its biology? As our understanding of the biology of *M. ulcerans* improves, this criterion will be answered.

Given this framework, how likely is it that Naucoridae and/or Belostomatidae are vectors of the Buruli ulcer disease? Herein, we explore the correlation in time and space between the proposed vectors, Naucoridae and Belostomatidae, and the Buruli ulcer prevalence. We discuss the other Bradford-Hill criteria, but do not focus on them specifically, as it was not directly within the scope of this work. Based on sampling in Cameroon, we characterise the set of suitable habitats within which species of the Families Belostomatidae and Naucoridae can maintain a population (their ecological niche) and describe the spatial distribution of these suitable habitats across West Africa. We then

explore any correlations between habitat suitability and Buruli ulcer prevalence at multiple spatial scales.

## Materials and Methods

### *Distribution of Belostomatidae and Naucoridae aquatic insect families*

Data were collected as described in S. Meyin à Ebong *et al.* (2012), hereafter referred to as the SME dataset. In brief, 36 sample sites in Cameroon were visited monthly from September 2012 to February 2013 (Figure 30), a period including both wet and dry seasons. Dip net sampling was conducted at all sites. Due to limitations of current taxonomic keys, the aquatic insects of interest were only identifiable to the phylogenetic division of Family. A second dataset was used in model validation, using data collected separately by A. Garchitorena (Figure 30). This dataset was collected as described in Garchitorena *et al.* (2014), and is hereafter referred to as the AG dataset. Niche models of Belostomatidae and Naucoridae are constructed using the SME dataset, and then tested on their ability to reproduce the independent AG dataset.

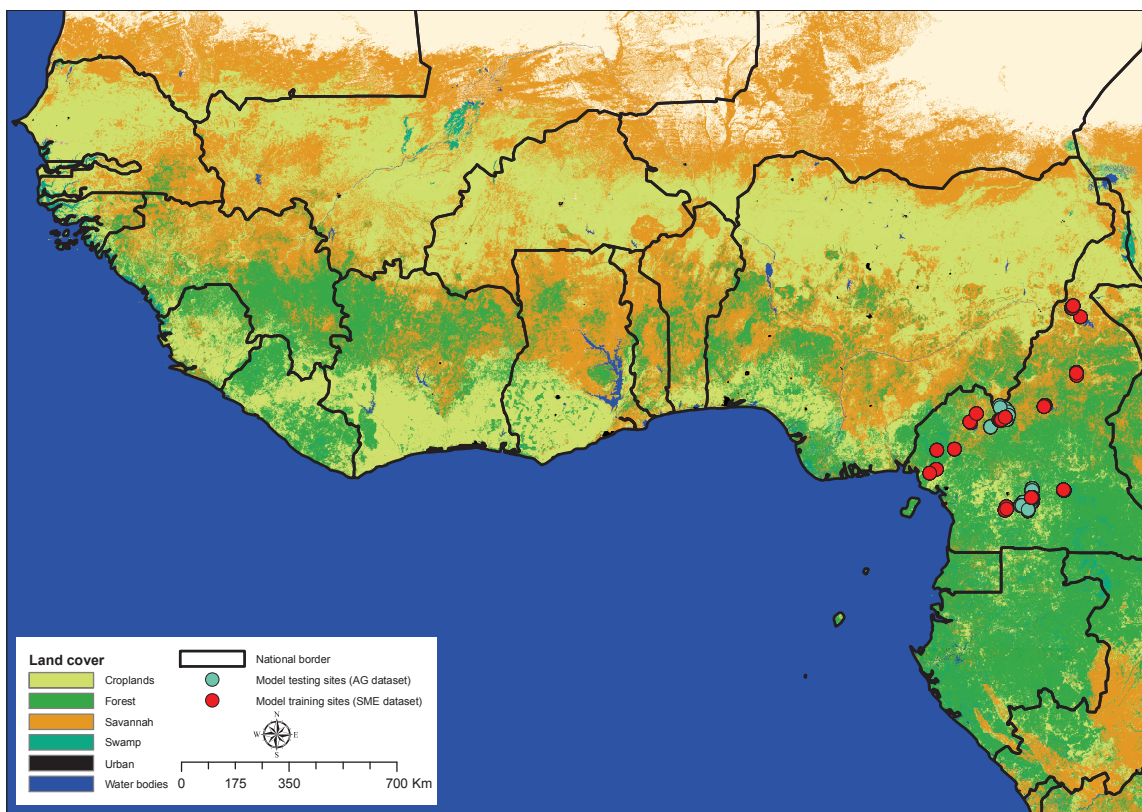


Figure 30. Study sites Cameroon against local land cover. Data from the SME dataset is in red, and AG data set is in blue. The sample sites span the extent of Cameroon, sampling from every major land cover category.

### *Environmental parameters used in ecological niche modelling*

Five ecological parameters were used to describe the distribution of suitable habitats: rainfall, flow accumulation, wetness index, land cover at the sample site, and land cover within the flight range of adult insects. These variables were selected on the basis of their likelihood to influence the distribution and condition of water, and are summarised in Table 8. Rainfall was highly seasonal, so we divide models by the season of collection. Models constructed using species distribution data from the dry season used the precipitation in the driest season; models constructed using species distribution data from the wet season used the precipitation in the wettest season. These two variables were taken from the Worldclim database, as BIO13 Precipitation of Wettest Month and BIO14 Precipitation of Driest Month (Hijmans *et al.* 2005). Flow accumulation was derived using elevation data (Jarvis *et al.* 2008). Flow accumulation is the surface area contributing water to a particular point, and indicates the potential amount of water available, which is then determined by rainfall. Using the SRTM elevation, flow accumulation was derived using the Fill, Flow Direction and Flow Accumulation tools in ArcMap 10.1 (ESRI). Wetness index has previously been shown to be associated with the Buruli ulcer (Wagner *et al.* 2008). In ecological terms it indicates the topological potential for water to accumulate, and was derived according to Equation 2,

$$WI = \log\left(\frac{FA * 500}{\tan(S)}\right)$$

Equation 2

where *FA* was the flow accumulation, 500 was the cell size in meters, and *S* was the surface slope in degrees. Large flow accumulation values and flat slopes resulted in high wetness index values, and indicate areas where water is likely to stagnate. In areas where the slope is zero wetness index had no value. Land cover was derived from the Global Land Cover Map 2009. GLC-At Point (hereafter GLC-AP) is the land cover at the sample site. In a 5km radius around the site the most common (modal) land cover category was described. For example, a sample site may be in savannah, but surrounded by forest. 5km was selected as the approximate flight radius of the insects (Bowden 1964, Robertson 1976, Lytle 1999, Mukai and Ishii 2007). This is termed GLC-5K.

Table 8. Environmental variables used in ecological niche modelling. All data were resampled to a spatial resolution of 0.004 decimal degrees (~300m<sup>2</sup>) for use in Maxent. Resampling used *resample()* in the library 'raster' of the software R (R core development team 2013).

Variable	Units	Original resolution	Source
Wetness index	m <sup>2</sup>	15 arc-sec (approx 450m <sup>2</sup> )	SRTM
Flow accumulation	m <sup>2</sup>	15 arc-sec (approx 450m <sup>2</sup> )	SRTM
Precipitation in wettest season	millimeter	30 arc-sec (approx 1 km <sup>2</sup> )	Bioclim 13
Precipitation in driest season	millimeter	30 arc-sec (approx 1 km <sup>2</sup> )	Bioclim 14
GLC-AP	Unitless	300m <sup>2</sup>	GLC 2009
GLC-5K	Unitless	300m <sup>2</sup>	GLC 2009

*Data preparation, niche modelling, and prediction of spatial distribution of suitable habitat.*

Insect distribution data from the 36 sample sites (Figure 30) were explored to identify normality, homogeneity of variance and correlation in the five environmental parameters used. Two sites were excluded from analysis as apparent outliers in flow accumulation, otherwise the data were normally distributed and homogenous. Correlation was observed between flow accumulation and wetness index and between precipitation in the driest and in the wettest seasons, however this was not significant ( $p > 0.05$ ) using a Spearman's correlation test. All data were resampled to a spatial resolution of 0.004 decimal degrees (~300m<sup>2</sup>) for modelling using *resample()* in the library 'raster' of the software R (R core development team 2013).

Across the scale of West Africa we assume that absence data is not reliable, as it is more likely to indicate failure of detection rather than evidence of absence. For this reason, we chose to conduct presence-only modelling, and the specific method selected was Maximum Entropy (Phillips *et al.* 2006). We used the software Maxent (Maxent 3.3.3k, Phillips *et al.* 2004) to construct these models. Maxent has been used several times in the past to model the ecological niche of disease vectors (González *et al.* 2010). Maximum entropy modelling minimises the divergence between the distribution of the environmental parameters and the species distribution, assuming the species is distributed in the most ecologically efficient manner possible. Maxent describes the environmental parameters across the study region, and is sensitive to the size of the study region. Herein, we generated background points in every raster cell within the extent of our study (that is, every environment was described). The models were replicated 100 times and averaged.

In order to select an appropriate extent to the area considered for modelling we chose to confine the spatial extent of our study to the region ecologically interpolated by the sample sites (Figure 31). The sample sites of the SME dataset describe a particular range of environmental conditions for the



five environmental parameters (Table 8). We excluded conditions that were not studied by the SME regime (much larger or lower values of rainfall, land cover categories not sampled, flow accumulation not observed at the study sites) to avoid extrapolating beyond the range of ecological parameter values studied. To select this interpolative region we constructed a Multivariate Environmental Similarity Surface (MESS map, Elith *et al.* 2010), excluding areas with negative MESS values from our study region. A MESS map was constructed for each model; models with more samples can be interpolated into larger ecological regions. MESS maps were constructed with the function *mess()* available in the library *dismo* in R.

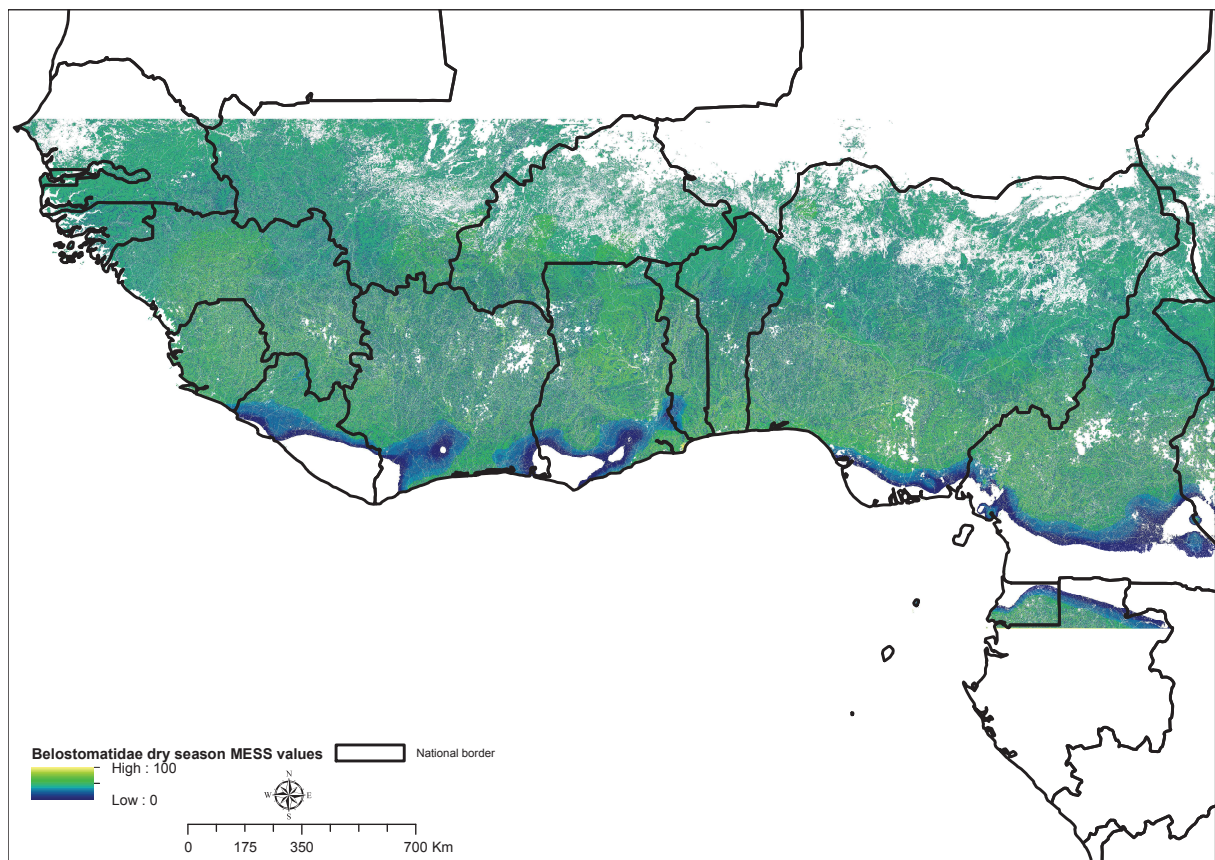


Figure 31. Delineation of the study area. We exclude areas of ecological extrapolation. Across West Africa the blue-green coloured region is ecologically interpolative within our sites, according to a multivariate environmental similarity surface (MESS). Regions where we extrapolate beyond the ecology of our study sites are identified using MESS values less than 0. Colour values are MESS values, indicating similarity to sample sites; a value of 100 is identical to the median of the sample sites. The spatial projection of this interpolative region is shown against a map of country borders of West Africa, for Belostomatidae adults in the dry season. This projection changes for each family, developmental stage and season, and though it is spatially discontinuous it is ecologically homogenous.

### *Evaluating model performance*

We evaluated the performance of these models according to their ability to correctly predict the distribution of the insects (with a separate dataset, the AG dataset) as efficiently as possible (avoiding overfitting). We used two methods to evaluate the models. First, the Akaike information criterion, corrected for small sample size (AICc), was used. AICc was generated as in Warren and Seifert (2011), and penalises complex models to avoid over-fitting the data, low AICc scores indicate the model is not over-fitting. Second, we tested the models ability to predict the AG dataset. The Area Under Curve (AUC) is often used to evaluate Maxent models, and has been criticized previously (Lobo *et al.* 2008). We used a modified version of the AUC for model evaluation, termed here AG AUC. The model predictions were compared to the known values as collected by AG. AG AUC values range from 0 to 1, values close to 1 indicate good performance, 0.5 is no better than random. Use of the AG dataset allowed us to use true absence data for model validation, avoiding the problem of pseudo-absence data in Maxent.

The purpose of these two different metrics is to consider different aspects of performance, neither were without limitations. Use of the AG dataset allowed a degree of validation across methodologies, indicating the extent to which our results were dependent on a particular sampling regime. AICc is traditionally used to indicate over-fitting in models, however it is sensitive to the size of the ecological niche of the species.

### *Identifying the relationship between habitat suitability and Buruli ulcer prevalence.*

Buruli ulcer prevalence data was collected for two endemic regions in Cameroon, Akonolinga (Landier *et al.* 2014) and Bankim (Marion *et al.* 2011), as shown in Figure 32. Around the centre of each village a buffer was created, and average habitat suitability in this buffer was correlated to the village Buruli ulcer prevalence using Spearman's rank correlation coefficient. Seven buffers were used to explore the effect of buffer size and shape. Around the centre of the village circular buffers of 1, 2, 3, 4, 5 and 10 km were selected, and average habitat suitability recorded. We also used a buffer defined by the borders of the village (Figure 32). This buffer changes in size for each village, and represents the approximate extent of the village area. 5km is approximately the flight radius of the insects, and the distance easily walkable by local people on an average day in Akonolinga (Giles-Vernick *et al.* 2014). To deal with this multiple testing, Bonferonni's correction of the *p*-values was used.

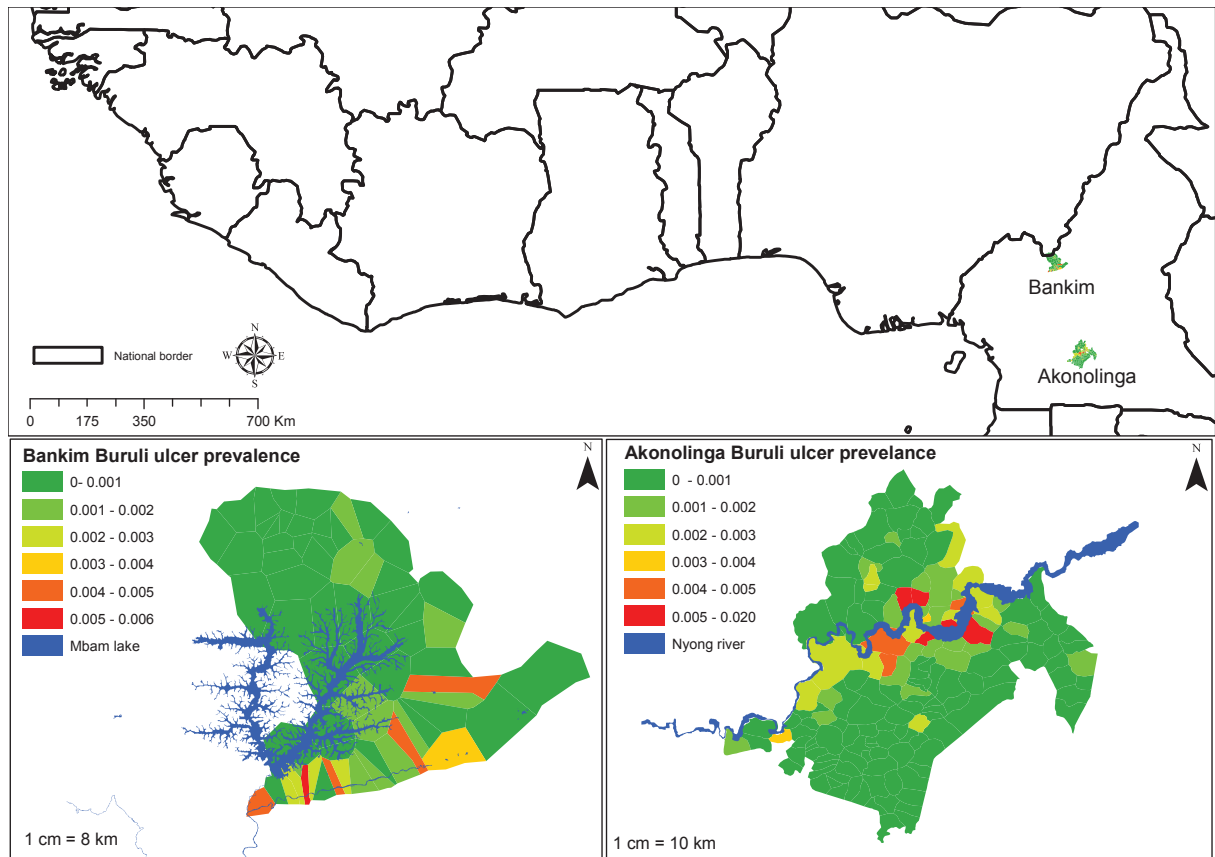


Figure 32. Spatial distribution of Buruli ulcer prevalence in Akonolianga and Bankim, two endemic regions within Cameroon.

## Results

### *Distribution of suitable habitat, and its relationship to Buruli ulcer*

In Akonolinga there was a significantly positive correlation between Buruli ulcer prevalence and average habitat suitability, for both Naucoridae and Belostomatidae, in the wet season (Table 9, Figure 33). This relationship was significant at multiple buffer distances. In contrast, in Bankim there was no significant correlation between Buruli ulcer prevalence and Belostomatidae or Naucoridae average habitat suitability, in either wet or dry seasons or at any buffer distance (Table 10).

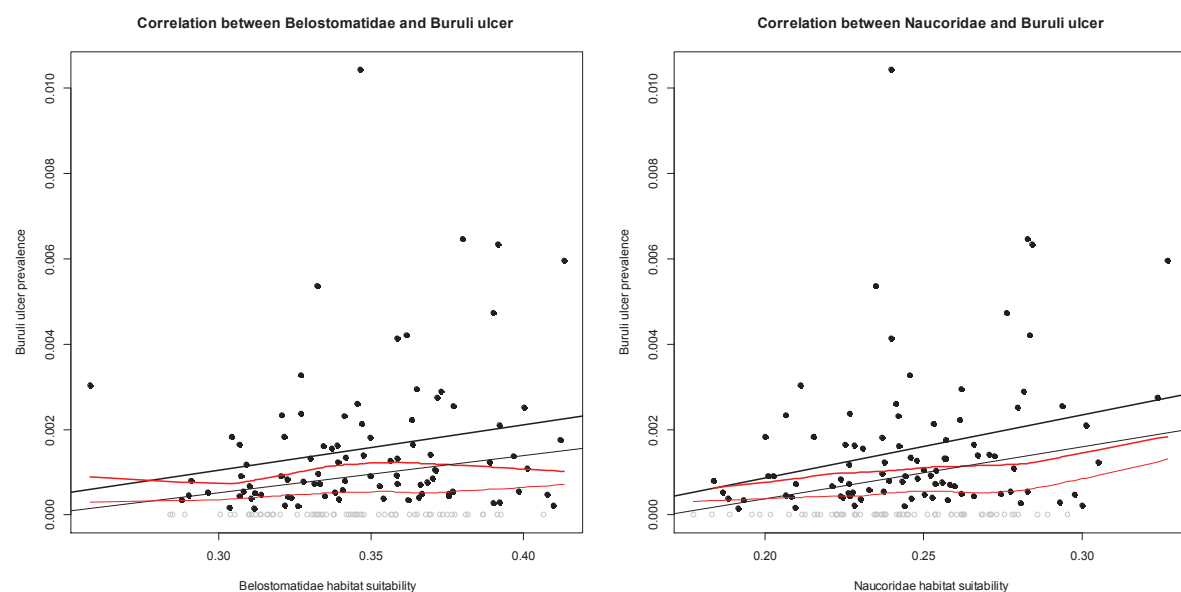


Figure 33. Correlation between the prevalence of Buruli ulcer and habitat suitability of Belostomatidae (Left) and Naucoridae (Right) in the wet season in Akonolinga. Colour indicates use of a linear model (Black) or locally weighted scatterplot smoothing (Red), different ways of viewing the correlation. Buruli ulcer was absent from certain villages (grey dots) where habitat suitability for the insects is high. Because this can skew any correlation between habitat suitability and prevalence, we explored the effect of including (thin lines) or excluding (thick lines) these villages. We note that in either case Spearman's rank correlation coefficient was significant. For Belostomatidae with Buruli ulcer absent villages  $p=0.08$ , without BU absent villages  $p=0.03$ . For Naucoridae with BU absent villages  $p=0.04$ , without BU absent villages  $p=0.02$ .

Table 9. Spearman’s rank correlation coefficients for correlation between Buruli ulcer prevalence and habitat suitability in Akonolinga, for both seasons and species. The column labelled Buffer is the distance, in km, around the village centre that the habitat suitability is considered, the buffer labelled village uses village boarders as a buffer (Figure 32). Bonferonni’s *p* value is the significance of the correlation between the insect and the disease, for clarity only significant values (<0.05) are presented, non-significant values are marked “-”. Significant positive correlations were observed between both Belostomatidae and Naucoridae in the wet season and Buruli ulcer prevalence, but not the dry season.

Species	Season	Buffer	Bonferroni <i>p</i> value
Belostomatidae	dry	10	-
		5	-
		4	-
		3	-
		2	-
		1	-
		village	-
	wet	10	0.000
		5	0.014
		4	0.023
		3	-
		2	-
		1	-
village		-	
Naucoridae	dry	10	-
		5	-
		4	-
		3	-
		2	-
		1	-
		village	-
	wet	10	0.001
		5	0.013
		4	0.014
		3	-
		2	-
		1	-
village		-	

Table 10. Spearman rank correlation coefficients for correlation between Buruli ulcer prevalence and habitat suitability in Bankim, for both seasons and species. The column labelled Buffer is the distance, in km, around the village centre that the habitat suitability is considered, the buffer labelled village uses village borders as a buffer (Figure 32). Bonferonni's *p* value is the significance of the correlation between the insect and the disease, for clarity only significant values (<0.05) are presented, non-significant values are marked "-". No significant correlations were observed in Bankim.

Species	Season	Buffer	Bonferroni <i>p</i> value
Belostomatidae	dry	10	-
		5	-
		4	-
		3	-
		2	-
		1	-
		village	-
	wet	10	-
		5	-
		4	-
		3	-
		2	-
		1	-
		village	-
Naucoridae	dry	10	-
		5	-
		4	-
		3	-
		2	-
		1	-
		village	-
	wet	10	-
		5	-
		4	-
		3	-
		2	-
		1	-
		village	-

### Ecologically important variables in the distribution of the aquatic insect families

Variable importance was evaluated using Jackknife variable removal. Jackknife removes a variable and evaluates the effect of variable removal on the model. In the dry season Belostomatidae and Naucoridae responded in broadly similar fashions; the variable whose removal had the largest effect was GLC 5km (Figure 34). The land cover categories most suitable for both Belostomatidae and Naucoridae are water bodies, artificial areas, rain fed croplands and forest/grassland mosaics (Figure 35). If one of these categories is the dominant category in 5km radius, in the dry season, the likelihood of encountering the insect is higher. Unsuitable categories were forest and vegetation/cropland mosaic.

In the wet season precipitation is more important than land cover. Precipitation suitability peaks at approximately 300 millimeters per month, and diminishes above or below this (Figure 35). For the dry season there is a simple increase in habitat suitability with increasing precipitation.

Flow accumulation had a negative association with habitat suitability, and wetness index had a positive association, regardless of season, for both Belostomatidae and Naucoridae.

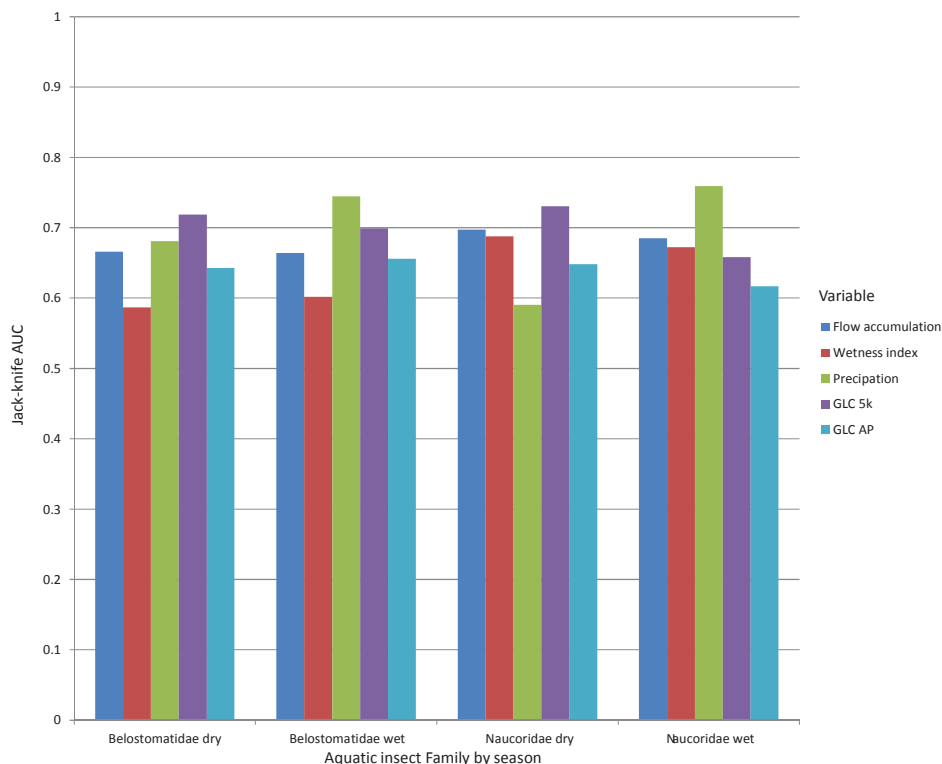


Figure 34. Importance of each variable according to Jack-knife AUC for wet and dry seasons. A high value indicates the variable is important; however this is sensitive to correlation within the variables. For both insects the most important variable in the dry season was the land cover in the flight radius (GLC 5km), in the wet season precipitation was the most important variable.

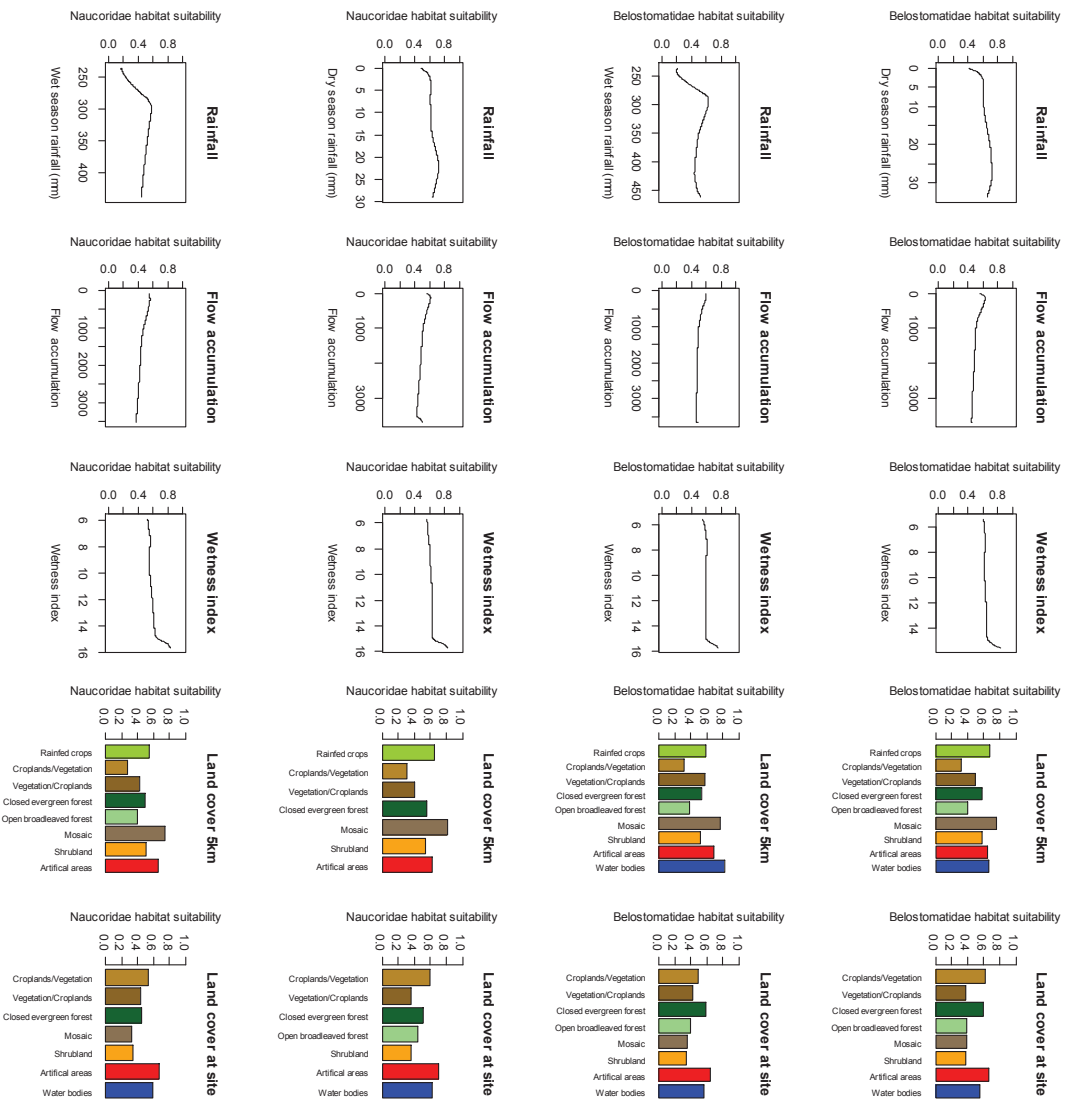


Figure 35. Habitat suitability for Belostomatidae (1<sup>st</sup> and 2<sup>nd</sup> rows) and Naucoridae (3<sup>rd</sup> and 4<sup>th</sup> rows) in the wet and dry seasons. Both insects have a negative relationship with flow accumulation, and a positive relationship with wetness index; an unexpected relationship given that increasing flow accumulation normally means increasing wetness index.



### *Model performance*

AICc for Naucoridae adults (dry season) was 14.6, and 14.2 in the wet season. For Belostomatidae adults (dry season) the AICc was 12.5, and 12.2 in the wet season. Scores of overfitting are relative; these scores indicate the Belostomatidae model was less prone to overfitting than the Naucoridae model.

The AG data set was also used in model validation. In the dry season Naucoridae adults had an AG AUC of 0.83, and 0.80 in the wet season. Belostomatidae adults had an AG AUC of 0.80 in the dry season, and 0.86 in the wet season. These scores indicate that the models are able to describe the distribution of the insects with good accuracy; the model based on SME dataset is able to accurately replicate the independently collected AG dataset.

### **Discussion**

We explored the correlation between the distribution of Belostomatidae and Naucoridae and the prevalence of Buruli ulcer. We have found a positive gradient between habitat suitability of Naucoridae and Belostomatidae and Buruli ulcer prevalence. Correlation does not imply causation; this result is not proof that the insects are vectors, however the existence of this correlation at certain times and locations argues for the possibility of a locally important transmission route.

There are significant temporal changes in this correlation between habitat suitability of the insects and Buruli ulcer prevalence; in Akonolinga the Buruli ulcer prevalence is correlated to Naucoridae and Belostomatidae distribution in the wet season, but not in the dry season. Buruli ulcer is known to have complex temporal changes in prevalence (Morris *et al.* 2014, Landier *et al.* 2014, van Ravensway *et al.* 2012), as is *M. ulcerans* (Marion *et al.* 2010, McIntosh *et al.* 2014). It is therefore unsurprising that, if these insects are implicated in transmission in some way (either as key hosts or vectors), the correlation between Buruli ulcer prevalence and their abundance would change in time. We also observe spatial changes in the correlation; water bug habitat suitability is not correlated to Buruli ulcer prevalence in Bankim, 457km North of Akonolinga. Speculatively, perhaps other routes of transmission may be more important in this region, for example contact with infected plant biofilms, as shown for Ghana (Merritt *et al.* 2005).

How do we interpret this result in terms of the Bradford-Hill guidelines? Herein we have focused on the correlation between these insects and the prevalence of the disease in both space and time, and while there is a significant positive correlation for the predicted abundance of the aquatic insects and the prevalence of the Buruli ulcer, this correlation is not consistent from region to region.

Previous research has proposed that *M. ulcerans* is transmitted from its environmental reservoir to humans by multiple modes of transmission (Roche *et al.* 2013). In such a situation of multiple vectors the relative importance of any given mode of transmission may be expected to vary in time or space, and our results are consistent with, though not proof of, this hypothesis. The lack of a clear signal between water bugs and Buruli ulcer in Bankim would suggest that other hosts or vectors may, perhaps, be key reservoirs or vectors in that region.

The ecological niches of both Naucorid and Belostomatid water bugs in West Africa are predominantly determined by the distribution of suitable landcover in a 5km radius, preferring water bodies, artificial areas and rain fed croplands. The specific land cover at the point of the site (GLC-AP) was less informative. The observation that the most suitable regional land cover class is water bodies is not surprising, but the high suitability of urban areas is curious. Ecologically this could have a variety of causes; there may be changes in the chemical composition of water in these habitats, a reduction in predation pressure, or a greater abundance of food. The specific reasons will require further research.

Our study has been limited in certain points; the low taxonomic resolution of the insects is a current limitation in this study. Secondly, an important limitation is that the distribution of *M. ulcerans* in these insects in these areas is unknown. The distribution of Naucoridae and Belostomatidae infected by *M. ulcerans* may differ from the distribution of Naucoridae and Belostomatidae generally. However, the insects are known to host the bacillus on their carapace, in their body (Benbow *et al.* 2008, Marion *et al.* 2010, Garchitorena *et al.* 2014) and in their salivary glands (Marion *et al.* 2010) in the wild, and the distribution of the insect necessarily sets a limit to the distribution of infected insects. A related limitation is the unknown incubation time of *M. ulcerans*; the time from infection to presentation at the hospital, remains unknown. Finally, we have only addressed a single criterion of the Bradford-Hill guidelines; correlation. We have not aimed for a full discussion of the other criteria, and our findings should not be interpreted as proof of the role of these insects as vectors or hosts. Rather, we have discussed the existence of, and change in, a correlation between these insects and Buruli ulcer. Future work aims to explore spatial variation in the correlation between Buruli ulcer and the entire plant and animal communities, identifying any similarities between regions where the correlation exists.

Despite these limitations, these results are consistent with previous research, which has shown that in Akonolinga the Nyong river is a risk factor for Buruli ulcer (Landier *et al.* 2014). Our results agree with this conclusion; the main focus of suitable habitats for the insects in Akonolinga is the Nyong

river, where the existence of large plants near the river banks provides appropriate habitat for Naucoridae and Belostomatidae to forage, develop and reproduce. Previous research has also implicated aquatic insects as important vectors in Akonolinga (Marion *et al.* 2010), including detection of *M. ulcerans* in the saliva of the insects.

In conclusion, we find a positive correlation between the abundance of Naucoridae and Belostomatidae suitable habitat and Buruli ulcer prevalence. This correlation is not constant, and changes in time and space. We interpret this as evidence consistent with that idea that Naucoridae and Belostomatidae act as locally important hosts or routes of transmission in certain conditions, which would be expected in the situation of multi-vectorial transmission.

## References

Amissah, A., *et al.* Investigating the Role of Free-living Amoebae as a Reservoir for *Mycobacterium ulcerans*. *PLoS neglected tropical diseases* 2014 8.9:e3148.

Benbow, E., *et al.* Aquatic invertebrates as unlikely vectors of Buruli ulcer disease. *Emerging infectious diseases* 2008 14.8:1247.

Elith, J., *et al.* The art of modelling range-shifting species. *Methods in ecology and evolution* 2010 1.4:330-342.

ESRI (Environmental Systems Resource Institute). 2009. ArcMap 10.1. ESRI, Redlands, California.

Garchitorena, A., *et al.* *Mycobacterium ulcerans* Ecological Dynamics and Its Association with Freshwater Ecosystems and Aquatic Communities: Results from a 12-Month Environmental Survey in Cameroon. *PLoS neglected tropical diseases* 2014 8.5:e2879.

Giles-Vernick, T., *et al.* The puzzle of Buruli ulcer transmission, ethno-ecological history and the end of “love” in the Akonolinga district, Cameroon. *Social Science & Medicine* (2014).

Global Land Cover Map 2009 USGS 2009, Global Land Survey, 2009, Landsat ETM+, 30m scene p102r062\_7dx19991002, USGS, Sioux Falls, South Dakota.

González, C., *et al.* Climate change and risk of leishmaniasis in North America: predictions from ecological niche models of vector and reservoir species. *PLoS neglected tropical diseases* 2010 4.1:e585.

Hill A., The environment and disease: association or causation? *Proc Roy Soc Med* 1965 58: 295–300.

Landier, J., *et al.* Space-time dynamics and landscape-associated risk of Buruli ulcer in Akonolinga, Cameroon *PLoS neglected tropical diseases* (2014) 8.9:e3123.

Marion, E., *et al.* Geographic expansion of Buruli ulcer disease, Cameroon. *Emerging infectious diseases* 2011 17.3:551.

Marion, E., *et al.* Seasonal and regional dynamics of *M. ulcerans* transmission in environmental context: deciphering the role of water bugs as hosts and vectors. *PLoS neglected tropical diseases* 2010 4.7:e731.

Marsollier, L., *et al.* Aquatic insects as a vector for *Mycobacterium ulcerans*. *Applied and environmental microbiology* 2002 68.9:4623-4628.

Marsollier, L., *et al.* Colonization of the salivary glands of *Naucoris cimicoides* by *Mycobacterium ulcerans* requires host plasmatocytes and a macrolide toxin, mycolactone. *Cellular microbiology* 2005 7.7:935-943.

Marsollier, L., *et al.* Impact of *Mycobacterium ulcerans* biofilm on transmissibility to ecological niches and Buruli ulcer pathogenesis. *PLoS pathogens* 2007 A 3.5:e62.

Marsollier, L., *et al.* Protection against *Mycobacterium ulcerans* lesion development by exposure to aquatic insect saliva. *PLoS medicine* 2007 B 4.2:e64.

Merritt, R., *et al.* Unraveling an emerging disease associated with disturbed aquatic environments: the case of Buruli ulcer. *Frontiers in Ecology and the Environment* 2005 3.6:323-331.

Merritt, R., *et al.* Ecology and transmission of Buruli ulcer disease: a systematic review. *PLoS neglected tropical diseases* 2010 4.12:e911.

Meyin A Ebong, S., *et al.* Survey of water bugs in Bankim, a new Buruli ulcer endemic area in Cameroon. *Journal of tropical medicine* 2012.

Morris, A., *et al.* Complex temporal climate signals drive the emergence of human water-borne disease *Emerging Microbes and Infections* 2014 3.8:e56.

Phillips, S., *et al.* A maximum entropy approach to species distribution modelling. *Proceedings of the 21st International Conference on Machine Learning* (2004). ACM Press, New York

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

Roche, B., *et al.* Identifying the Achilles heel of multi-host pathogens: the concept of keystone 'host' species illustrated by Mycobacterium ulcerans transmission. *Environmental Research Letters* 2013 8.4: 045009.

Schünemann, H., *et al.* The GRADE approach and Bradford Hill's criteria for causation. *Journal of epidemiology and community health* 2011 65.5:392-395

Jarvis, A., *et al.* Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database 2008 <http://srtm.csi.cgiar.org>.

Van der Werf, T., *et al.* Mycobacterium ulcerans infection. *The Lancet* 1999 354.9183:1013-1018.

van Ravensway, J., *et al.* Climate and landscape factors associated with Buruli ulcer incidence in Victoria, Australia. *PLoS one* 2012 7.12:e51074.

Veitch, M., *et al.* A large localized outbreak of Mycobacterium ulcerans infection on a temperate southern Australian island. *Epidemiology and infection* 1997 119.03:313-318.

WHO (2001) Buruli Ulcer: Fact Sheet No. 199 . World Health Organisation, Geneva

Wagner, T., *et al.* A landscape-based model for predicting Mycobacterium ulcerans infection (Buruli ulcer disease) presence in Benin, West Africa. *EcoHealth* 2008 5.1:69-79.

Warren, D., and Seifert, S. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecological Applications* 2011 21.2:335-342.

Hijmans, R., Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* 2005 25:1965-1978.

## Chapter 5: Synthesis

## Introduction

This thesis set out to describe the distribution of *M. ulcerans*, and the insects suspected to act as vectors of this pathogen, in Cameroon, Africa. *M. ulcerans* is an emerging infectious disease, afflicting more than 5,000 people per year, and knowledge of the distribution of the pathogen is needed in order to assist in control of the Buruli ulcer disease. At the beginning of this thesis I undertook a literature review of Buruli ulcer and *M. ulcerans*, with the aim of gaining more knowledge of this emerging pathogen. This was presented in Chapter 1, where the basic biology of *M. ulcerans* was introduced. I also undertook a review of different methods available for ecological niche modelling, which developed into Chapter 2, a discussion on the effect of study area and sample size in ecological niche modelling. The skills developed were applied practically to a description of the distribution of *M. ulcerans* in Akonolinga, Cameroon, (Chapter 3), and tested against a second database from French Guiana, South America. Finally, I described the distribution of the suspected vectors across West Africa (Chapter 4). The following is a synthesis of the previous three chapters, with a view towards future work and the key research questions these findings have identified.

### **Ecological niche models; sample size, surface area and performance.**

#### *Aim*

Following a review of methods available to construct ecological niche models, I undertook an evaluation of the effect of differences in sample size between different databases and the impact this had on model performance. As a part of the EXTRA-MU and LabEx CEBA collaboration networks I have engaged with researchers working in a variety of regions and with differing sample sizes, and the goal herein was to gain a better understanding of maximum entropy modelling, the Maxent software, and the importance of different scale and different sample size for model performance.

#### *Methods*

To explore these questions, I created a virtual species distributed according to elevation, precipitation and temperature. Samples were generated from this species to describe how the model responds to changes in the extent of the study and the sample sizes used. Following a review of the literature, I also explored the effect these changes have on a commonly used model evaluation metric, the Area Under the Curve (AUC).

## *Results*

The key findings were:

- 1) The AUC of a Maxent model increases with increasing study extent,
- 2) The AUC of a Maxent model increases with decreasing species niche volume,
- 3) The true performance of the model does not correlate to changes in the AUC.

As discussed in Chapter 2, the performance of the Maxent model, according to AUC, increases as the sample size increases (as expected), as the hypervolume,  $\mathbf{G}$ , of the study increases, and as the size of the ecological niche of the species,  $V$ , decreases.

However, this increase is not a real increase, and is a result of the generation of pseudo-absence points in the evaluation metric. As ratio of  $V$  to  $\mathbf{G}$ , of volume of the entire set of environments under consideration to the volume of that set occupied by the virtual species, approaches 0, the AUC approaches 1. The pseudo-absence points are more likely to be in an absent region.

### *Implications of these results within the research community of epidemiologists and disease ecologists*

The results of the research are presented in the paper “How many samples do we need to describe the spatial distribution of an emerging pathogen?”, which at the time this thesis was submitted is being prepared for publication. These findings were presented in a manner accessible to epidemiologists, as, though these ideas are not unknown in ecology (Warren and Seifert 2011), they are less widely appreciated in disease epidemiology. The need to conduct sampling based on ecological principles has been a key point of communication within the research network. As is shown in the Chapter, inappropriate sampling can dramatically bias our findings. This work was applied to *M. ulcerans* and Buruli ulcer in the following two papers.

### *Future work*

In this chapter I had not explored the importance of sample bias. Often sampling methodologies focus more effort on one geographic region, translating into a better understanding of the ecological niche under the environmental parameters that prevail in that region. Often this bias is an unavoidable result of sampling, for example sample sites are often near roads as they must be accessible. Future work on this subject aims to explore techniques apart from the AUC, and the impact of sampling bias on sample size needed for niche modelling. Given the difficulty in finding available data for emerging infectious diseases, a current research priority is the efficient use of the data available.



## The ecological niche of *Mycobacterium ulcerans*

### *Aim*

The original aim of the thesis, as outlined within the EXTRA-MU project, was to describe the distribution of *M. ulcerans* in Akonolinga, Cameroon. After undertaking a literature review of *M. ulcerans*, the importance of land cover (Wagner *et al.* 2008) and season (van Ravensway *et al.* 2012, Morris *et al.* 2014B) became apparent. These variables were included in the models, along with topography, which influences the flow and accumulation of water. The models were then interpolated across Akonolinga to generate the first hazard maps of the distribution of the pathogen for this Buruli ulcer endemic region.

### *Methods*

Following the review of available methods, I undertook general linear modelling to describe the distribution *M. ulcerans* in Akonolinga. Sampling was biotic, taking the 5 dominant insect Orders of the local community (normally making up 80-90% of the biotic sample), and described the distribution of infected insects in relation to the land cover and topography of the sample site.

Land cover and topography were described within two buffers around the sample sites, a local buffer of 5km (approximately the flight radius of adults of the 5 Orders), and watershed buffers (the drainage area of the sample site).

### *Results*

The key findings were:

- 1) Retention of both the watershed and 5km buffers as important factors driving *M. ulcerans* distribution,
- 2) A seasonal distribution,
- 3) Importance of lowland swamp and low amounts of urban habitat,
- 4) Sensitivity to spatial extent,
- 5) The distribution of *M. ulcerans* is predicted to be generalist.

The finding that both the watershed buffer, and the 5km buffer, are important and retained in model is interpreted as signifying an important non-aquatic component to the distribution of *M. ulcerans*. This is not biologically unrealistic, as in Australia the pathogen is commonly found in non-

aquatic environments, such as urban areas (Carson *et al.* 2014), but a great deal of sampling in Africa is aquatic. This argues for the need for broader, non-aquatic sampling of the bacterium.

The seasonal distribution of the bacterium may be expected given the notably seasonal distribution of the disease (van Ravensway *et al.* 2012, Morris *et al.* 2014B). In the wet season, in Akonolinga, *M. ulcerans* is less common in small watersheds that drain wet swamps, while in the dry season the bacterium is less common in large watersheds that drain flood plains. It has been speculated that the seasonal change in the disease may be due to changes in either human behaviour or vector abundance (Johnson *et al.* 2005), however these findings support the importance of changing distribution of the bacterium as a potential cause of the change in prevalence of Buruli ulcer, however the unknown incubation time of the bacterium remains as a limitation to this interpretation.

In collaboration with researchers at the LabEx CEBA, I have also demonstrated the sensitivity of the Akonolinga model to extrapolation beyond the spatial extent of the study area, as it is unable to predict the distribution of the pathogen in French Guiana. This is consistent with current ideas of the importance of various ecovars of *M. ulcerans*; the species has recently undergone a population bottleneck and if it is still adapting to a new set of environments we would expect that local adaptation may cause notable changes in the population distribution. This will have consequences for future detection of the bacterium, given that previous work has used only the IS2404 probe for PCR, which is now being recognised as non-specific (Tobias *et al.* 2013); however based on these findings I would expect different ecovars to reveal different ecological distributions. This may aid in understanding the method of transmission, as (speculatively) not all ecovars may be equally pathogenic towards humans.

### *Limitations*

This research has suffered certain limitations. *M. ulcerans* detection was with IS2404 and KR probes, this is considered a strong indication of the presence of *M. ulcerans*, but does not differentiate between ecovars of the species, between pathogenic and non-pathogenic variants of the bacterium, or between viable and inviable bacteria. The hazard distribution maps can therefore be considered as maps of the potential distribution of the bacterium, without finer phylogenetic resolution.

While I described the distribution of *M. ulcerans* at the landscape scale, the bacterium could also be described at larger or smaller spatial scales, and the scale will influence the outcome of the model (as discussed in Chapter 2). The choice of the landscape scale was a practical decision, as at this scale

we are able to generate maps of the distribution of the bacterium that can be used by local health officials. My results cannot describe the distribution of *M. ulcerans* at larger or smaller scales, and the finding that the model does not predict from Akonolinga to French Guiana highlights the sensitivity of the models to not only scale, but also spatial extent.

The produced hazard maps are not risk maps. The maps of the predicted distribution of suitable habitat for *M. ulcerans* in Akonolinga describe an environmental hazard; in order for the hazard to be realised (for Buruli ulcer to occur) a suitable vector, host, reservoir and susceptible human population need to all co-occur at the same time. These hazard maps represent the first step towards generating maps of risk to the human population, as our understanding of the mode(s) of transmission improves.

Finally, a particular feature of our sample regime in Akonolinga is the use exclusively of biotic samples. Only the aquatic invertebrate community was considered, and the 5 most common Insect Orders were used to describe the distribution of *M. ulcerans*. Previous research has found the bacterium in the abiotic environment in French Guiana (Morris *et al.* 2014), and the bacterium is detected in water filtrand and soils in Ghana (Williamson *et al.* 2012), the results are therefore applicable only to this biotic component of the distribution. However, the biotic component does appear to be important, as the insects are persistently and commonly infected (Mosi *et al.* 2008).

### *Policy Implications*

The apparently generalist distribution of *M. ulcerans* in Akonolinga has implications for control of the disease, depending on our interpretation occupancy of the suitable habitat.

Firstly, it is possible the habitat is suitable, but currently uninhabited (as discussed in Chapter 1). This is commonly the case with invasive species, where a section of the habitat will soon be exploited. Given the previously discussed correlation between the distribution of the bacterium and the disease (Williamson *et al.* 2008), and the observation of a progression of disease prevalence, in space and time, through Akonolinga (Landier *et al.* 2014), and the finding presented herein (that suitable habitat is common), implies that *M. ulcerans* may be expanding through Akonolinga. If this is so, it would be expected to continue expanding, likely resulting in an increase in Buruli ulcer prevalence in the coming years.

However, there is an alternative interpretation. These results find that there is a highly generalist distribution of suitable habitat. This suitable habitat may be entirely occupied by the bacterium, in which case the historical expansion of the disease may be due to either historical changes in human

behaviour (Giles-Vernick *et al.* 2014), in the land use of local communities, or changes in the distribution of key vector(s) or host(s). In such a case *M. ulcerans* may be common in the environment, but only able to cause human infection under certain circumstances.

In either interpretation, methods of controlling the Buruli ulcer which aim to remove the bacterium from the environment are unlikely to be cost-effective, as a large amount of suitable habitat would have to be made unsuitable for the bacterium. Control methods aimed at altering human behaviour may be more successful in the alternative interpretation, that the bacterium occupies all of its suitable habitat. The importance of distinguishing between these two proposed hypotheses is discussed further below, in the section on future research avenues.

The results of the research are presented in the paper “Topography and land cover of watersheds predicts the distribution of the environmental pathogen *Mycobacterium ulcerans* in aquatic insects.”, which at the time this thesis was submitted had been editorially accepted for publication in PLoS Neglected Tropical Diseases.

#### *Future work*

I described the distribution of *M. ulcerans* in the 5 dominant insect Orders in the region, given the importance of biotic interactions for the species (Roche *et al.* 2013) future work aims to explore this again with a finer level of detail, at the Family level. Current work has shown the sensitivity to extrapolation, collaboration with field researchers in Benin is expected to yield an additional area of study, for comparison with a third region. Understanding the sensitivity to spatial extrapolation and how this changes from region to region should yield further information on the environmental drivers of *M. ulcerans* distribution.

## **The relationship between Belostomatidae, Naucoridae and Buruli ulcer.**

### *Aim*

Review of the literature on *M. ulcerans* and Buruli ulcer reveals the ongoing debate on the mode of transmission of *M. ulcerans* from the environment to humans (Marsollier *et al.* 2002, Benbow *et al.* 2008). I aimed to contribute to this discussion, through exploration of any correlation between the predicted abundance of Belostomatidae or Naucoridae with Buruli ulcer prevalence. To contribute in a constructive manner, I employed the Bradford-Hill guidelines for vector identification (Hill 1965). These guidelines lay a set of criteria that an insect species suspected to be a vector of a disease would be expected to fulfil. These criteria are specificity, correlation, plausibility and coherence. I addressed the criterion of correlation, any vector should be correlated in abundance to the prevalence of the disease, and explored the distribution of Belostomatidae, Naucoridae and the Buruli ulcer for any correlations.

### *Methods*

Following collection of samples (Meyin à Ebong *et al.* 2012) and discussion on the ecology of Belostomatidae and Naucoridae, I selected maximum entropy modelling as the most suitable presence-only niche modelling method. As discussed in Chapter 2, this method is sensitive to the area of the study. I limited the extent of the study using multivariate environmental similarity surfaces, these enable the model to avoid extrapolating into un-sampled environments, and due to the extensive sampling regime (Meyin à Ebong *et al.* 2012) the models are able to interpolate into a large spatial area.

### *Results*

The key findings were:

- 1) Significant positive correlation between the insects and the disease,
- 2) Change in this correlation from season to season for certain regions,
- 3) Change in the correlation from region to region (within Buruli ulcer endemic regions),
- 4) Suitable habitat for the insects included urban habitats.

Vectors are normally correlated in abundance to the prevalence of the disease they transmit (Hill 1965, Merritt *et al.* 2010), and here I have found that these two water bug families are indeed positively correlated to the prevalence of Buruli ulcer. This is by no means proof, but is consistent with their potential role as vectors.

This correlation was found to be seasonal in nature, occurring in the Akonolinga wet season but not in the dry. This would suggest that the wet season is a key time in the epidemiology of Buruli ulcer transmission; however the incubation time is known to vary from region to region and among genders and age groups (Marion *et al.* in preparation), and the unknown incubation time in Akonolinga remains a limitation to this interpretation.

In a similar manner, the correlation changes from region to region. While there is a significant correlation in Akonolinga, there is no significant correlation between these insects and the Buruli ulcer prevalence in Bankim. I have taken care to explain the distinction between correlation and causation in Chapter 4, the fact that we find a correlation does not prove a causative role. However, in the context of the Bradford-Hill guidelines, I interpret the presence of, and change in, this correlation as consistent with the hypothesis of multi-host transmission.

Belostomatidae and Naucoridae were found to be positively correlated to urban habitats, the two families of insects appear to be more common near these urban areas. A potential limitation of this interpretation is, as discussed in Chapter 4, the possibility sampling bias favouring urban habitats. However, the dataset does not contain any obvious spatial correlation to urban habitats, so though this cannot be conclusively rejected, it appears unlikely. A similar limitation, discussed in Chapter 1, is detection bias. The sampling protocol may sample from all habitats equally, however the insects may be more detectable in certain habitats than in others. I have no evidence of this potential detection bias, and cannot comment on the likelihood of such an effect. If Belostomatidae and Naucoridae are indeed more common in urban habitat, barring sample or detection effects, it would lend evidence to their role as vectors. As discussed in Chapter 1, these insects rarely bite humans, and normally accidentally, so an increase in the contact rates between humans and the insects would increase this probabilistically low biting rate.

### *Limitations*

As discussed in the paper, we only describe the distribution of these two families, and do not discuss the distribution of other aquatic insects. If Belostomatidae and Naucoridae are vectors, or key hosts, they should be more strongly correlated to the distribution of the disease than other members of the community who are not vectors (Hill 1965). I aimed to address one aspect of the guidelines, and cannot comment on the other criteria.

The correlation is found to exist only in the wet season, suggesting that the wet season may be an important time in transmission (as discussed above). However, the time between infection by *M.*

*ulcerans* and patient presentation at a hospital remains unknown, and is suspected to be on the order of four to five months. Both the models for the distribution of *M. ulcerans* (Chapter 3) and the models for the distribution of the insects found seasonal changes in the distributions. These temporal changes are complex, and beyond the scope of this thesis. Future work aims to unravel more information on the incubation time of the infection, which should aid in understanding the complex seasonal drivers of the disease (Morris *et al.* 2014B, van Ravensway *et al.* 2012)

The models of Belostomatidae and Naucoridae distributions were constructed at the phylogenetic resolution of family. This has limitation was based on current taxonomic keys, and current work by Solange Meyin à Ebong is expected to address this, allowing more detailed exploration of the differences in the distribution of different species within the family (Meyin à Ebong *et al.*, forthcoming).

A related limitation for the article is the unknown distribution of *M. ulcerans* infected individuals of Belostomatidae and Naucoridae. The sampling regime at the time did not include the capacity to test the insects for the presence of *M. ulcerans*, potentially the distribution of the infected insects is different to the distribution of the insects in general. This will make an important component of future work.

The results of this research are presented in the paper “Ecological niche modelling of Hemipteran insects in Cameroon; the paradox of a vector-borne transmission for Mycobacterium ulcerans, the causative agent of Buruli ulcer”, which at the time this thesis was submitted had passed the first round of revision for the International Journal of Health Geographics. The findings presented provide empirical support for previous work by Roche *et al.* (2013), who discussed the view of *M. ulcerans* as a species within a trophic network, discussing the possibility of multiple vectors involved in transmission. While the finding of a spatially changing correlation is not proof of this hypothesis, it is consistent with it.

## Conclusion, practical applications, and questions for future research

Following the findings and limitations outlined above, I would argue for four key areas to be pursued in future research on Buruli ulcer:

- 1) Watershed based sampling protocols,
- 2) Integrating aquatic sampling with sampling of the non-aquatic community,
- 3) Evaluation of occupancy of suitable habitat in endemic regions,
- 4) Evaluation of spatial variation in the correlation between Buruli ulcer prevalence and aquatic insect abundance in West Africa.

Here I have found that watershed size and condition is an important factor in the distribution of the bacterium, and previous research has found that genetically distinct populations of *M. ulcerans* appear to be bound by watershed boundaries (Vandelannoote *et al.* 2014). These boundaries may present a migration barrier to the bacterium, preventing it from accessing certain areas, leading to genetic differences between populations. However, it appears unlikely that this barrier strictly prevents all movement; an important component of the bacterium's distribution appears to be described by the local conditions around the sample site. This would suggest the importance of trophic webs that can reach beyond aquatic environments (Roche *et al.* 2013). Such trophic interactions could enable the bacterium to cross this watershed boundary, and an understanding of the relative importance of these boundaries in different ecological contexts (different topography, different ecological communities, different trophic webs) may be key in explaining the distribution of the bacterium on a larger scale than those discussed herein.

The importance of these barriers to movement (*sensu* Figure 16, Chapter 2) could be evaluated with integrated aquatic and non-aquatic sampling. This would use sampling regimes based on the structure of the stream network and the position of the watershed boundaries. Figure 36 provides an example of this idea, though this Figure is strictly for the purpose of explaining the principle. A virtual sampling regime is presented, where two sampling methodologies are conducted simultaneously. Blue dots are aquatic samples, which describe the distribution of *M. ulcerans* aquatic sampling protocols along a stream network (as in Morris *et al.* 2014A). Given that my findings highlight the importance of the condition of the upstream area in predicting *M. ulcerans*, the following hypothesis can be proposed: if a sample is positive, the downstream area should be more likely to be positive for *M. ulcerans*. To complement this aquatic sampling, non-aquatic sampling protocols, for example of terrestrial plants and insects (green dots in Figure 36), run along a transect crossing the stream and watershed boundary. Distance to a stream is a risk factor for *M.*



*ulcerans* presence (Merritt *et al.* 2010); does the risk of presence change significantly when we cross the watershed boundary? Recent work in French Guiana discovered that, following construction of the Petit-Saut dam in Sinnamary, the number of cases downstream of the dam dropped dramatically (Morris *et al.* 2014A). This implies that changes to the watershed do dramatically change the distribution of Buruli ulcer, and the knowledge gained from integrated and ecologically justified sampling regimes should enable us to understand how these watershed changes alter *M. ulcerans* distribution, informing us on potential changes risks to human health.

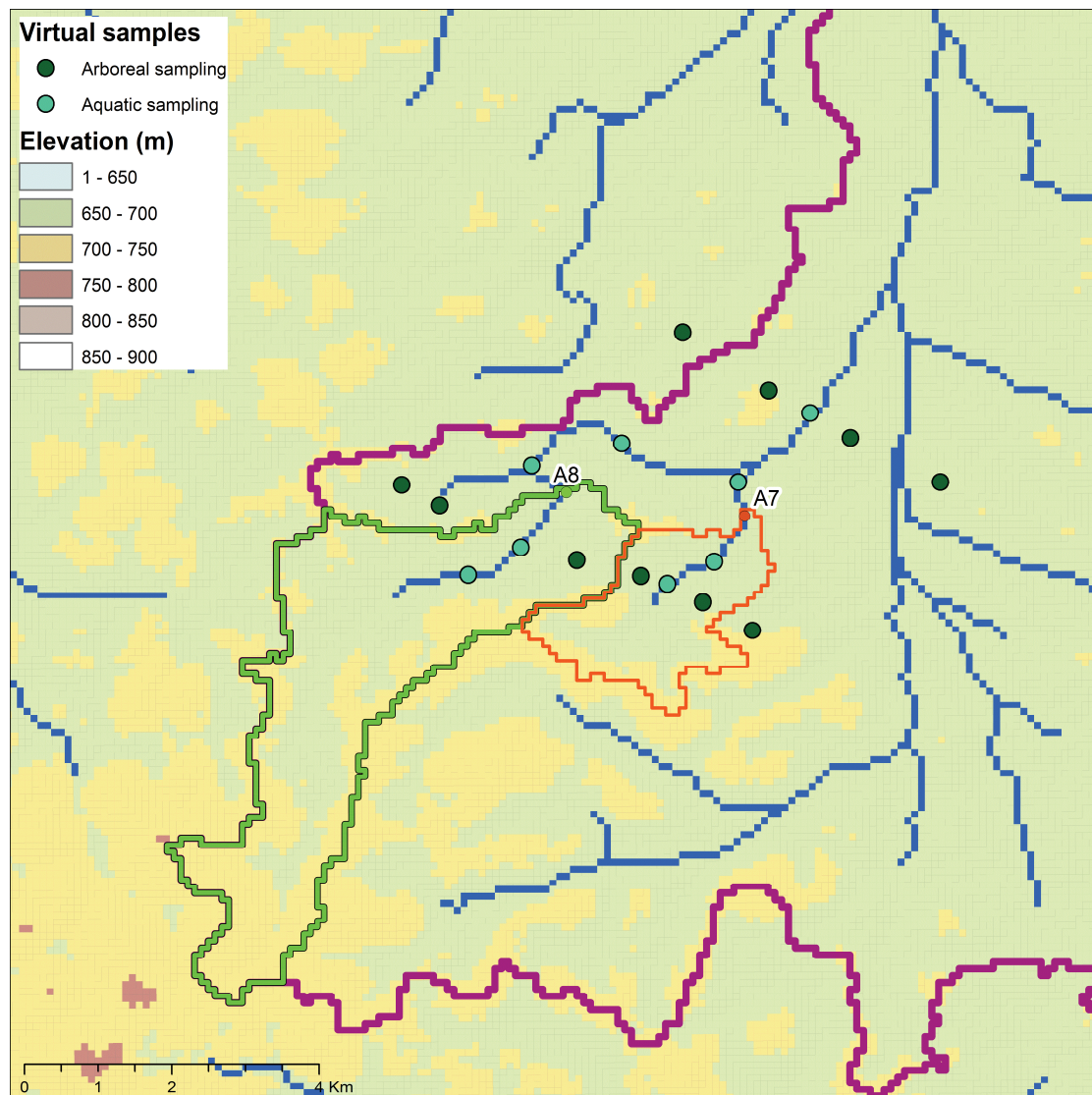


Figure 36. A potential sampling regime to explore the effect of watersheds as a boundary to *M. ulcerans*, here for Akonolinga, near sites A7 and A8 (Table 1). Three watershed boundaries are presented in orange, green and purple, and streams are in blue. Two sampling methodologies could be used, samples along the aquatic network (blue dots) and non-aquatic samples across the network (green dots) which transect the streams and watershed boundaries, collecting from the wider biotic network of plants and insects.

I have generated the first maps of the predicted distribution of suitable habitat for *M. ulcerans* at the landscape scale in Akonolinga. These maps describe the distribution of suitable habitat in the region, and predict that *M. ulcerans* is likely to be present throughout the region. Given the consequences to human health, the evaluation of the occupancy of this suitable habitat is a key area for research. The generated models perform well, and accurately describe the distribution of *M. ulcerans* at the sample sites, and on this basis the bacterium would be predicted to have a distribution that is generalist within Akonolinga. Such a prediction of generalist distribution of *M. ulcerans* opens two hypotheses; 1) that the habitat is suitable and uninhabited, or 2) that the habitat is suitable and inhabited. The fact that the bacterium appears to be common in the environment (Williamson *et al.* 2008, Garchitorena *et al.* 2014) lends support to the second hypothesis. However, identifying the proportion of suitable and occupied or unoccupied habitat will enable predictions about the future changes in the distribution of the bacterium.

The existence of the correlation between Belostomatidae and Naucoridae and Buruli ulcer prevalence supports, though does not prove, the idea that these insects are important vectors, in the context of the Bradford-Hill guidelines (Hill 1965). The correlation between their presence and the prevalence of the disease in Akonolinga would be expected if they are vectors, however there was no significant correlation found between their presence and the prevalence in Bankim. As discussed in Chapter 4, there are several potential reasons for this. Bankim is more recently endemic (based on Buruli ulcer case detection), and has undergone several unique changes in human distribution which may skew any relationship, different vectors may be locally important. Alternatively, the distribution of the disease prevalence and the insect abundance may overlap by chance in Akonolinga. I attempted to control for this with Bonferonni correction of the t-tests, however it is an important possibility that cannot be ruled out. A notable advantage of the extensive sampling regime of Meyin à Ebong *et al.* (2012) is the ability to interpolate over a large area. An important goal for future work is to explore the change in the correlation between Buruli ulcer and these insects in different regions of West Africa. Prevalence data can be collected from different administrative districts and used to test the hypothesis that there is some consistency in the sites where the correlation exists, for example perhaps there are similar trophic networks in areas where the correlation is significant. The identification of a consistent environmentally dependent correlation between these insects and the prevalence of the Buruli ulcer would support the multi-vectorial transmission hypothesis of this emerging neglected tropical disease.

## Conclusion

Starting from an evaluation of the ecology of *M. ulcerans* and a review of distribution modelling methods, I have produced ecological niche models of the pathogen and the suspected vectors, and gained extensive experience of the tools, techniques and challenges of niche modelling. The results of this experience are the hazard maps for the Buruli ulcer endemic region of Akonolinga, a new understanding of the non-aquatic ecology of the pathogen, and new arguments for the importance of Belostomatidae and Naucoridae in Buruli ulcer epidemiology. Future work will collaborate on research into the importance of genetic-ecological differences between difference ecovars of the species, developing protocols for integrated aquatic sampling methodologies, and analysis of larger scale spatial variation in the relationship between Belostomatidae and Naucoridae and Buruli ulcer.

## References

- Benbow, E., *et al.* (2008) Aquatic invertebrates as unlikely vectors of Buruli ulcer disease. *Emerging infectious diseases* 14.8: 1247.
- Garchitorena, A., *et al.* (2014) *Mycobacterium ulcerans* Ecological Dynamics and Its Association with Freshwater Ecosystems and Aquatic Communities: Results from a 12-Month Environmental Survey in Cameroon. *PLoS neglected tropical diseases* 8.5: e2879.
- Giles-Vernick, T, *et al.* (2014). The puzzle of Buruli ulcer transmission, ethno-ecological history and the end of “love” in the Akonolinga district, Cameroon. *Social Science & Medicine*.
- Hill A., (1965) The environment and disease: association or causation? *Proc Roy Soc Med* 58: 295–300.
- Johnson PDR, *et al.* (2005) Buruli Ulcer (*M. ulcerans* Infection): New Insights, New Hope for Disease Control. *PLoS Med* 2(5): e173. doi:10.1371/journal.pmed.0020173
- Landier, J., *et al.* Space-time dynamics and landscape-associated risk of Buruli ulcer in Akonolinga, Cameroon *PLoS neglected tropical diseases* (2014) 8.9:e3123.
- Meyin A Ebong, S., (2012) *et al.* Survey of water bugs in Bankim, a new Buruli ulcer endemic area in Cameroon. *Journal of tropical medicine*.
- Marion *et al.* (in preparation) Buruli ulcer in South Western Nigeria
- Marsollier, L., *et al.* (2002) Aquatic insects as a vector for *Mycobacterium ulcerans*. *Applied and environmental microbiology* 68.9: 4623-4628.
- Morris, A., *et al.* (2014 B) Complex temporal climate signals drive the emergence of human water-borne disease. *Emerging Microbes and Infections* 3, e56
- Tobias, N., *et al.* (2013) "Complete genome sequence of the frog pathogen *Mycobacterium ulcerans* ecovar Liflandii." *Journal of bacteriology* 195.3: 556-564.
- Vandelannoote, K, *et al.* (2014) Insertion Sequence Element Single Nucleotide Polymorphism Typing Provides Insights into the Population Structure and Evolution of *Mycobacterium ulcerans* across Africa. *Applied and environmental microbiology* 80.3: 1197-1209.
- van Ravensway, J, *et al.* (2012) Climate and landscape factors associated with Buruli ulcer incidence in Victoria, Australia. *PloS one* 7.12: e51074.
- Williamson, H., *et al.* (2008) "Distribution of *Mycobacterium ulcerans* in Buruli ulcer endemic and non-endemic aquatic sites in Ghana." *PLoS neglected tropical diseases* 2.3: e205.

## Chapter 6: Appendix

## Foreword

As a result of the collaboration network within the Extra-MU, BIOHOPSY and JEA1 AtoMyc projects part of this thesis has been devoted to the preparation of papers on which I am a co-author. Several of these papers are cited at certain points in the thesis, and in align with the requirements of the Ecole Doctorale SIBAGHE these are presented here, along with my contributions to each.

The papers are;

- 1) “Buruli ulcer in south western Nigeria Emerging infectious diseases.” (in preparation) E. Marion, **K. Carolan**, A. Adeye, M. Kempf, A. Chauty, L. Marsollier.
- 2) “Niche-based extinction of species boosts presence of the environmentally-acquired pathogen, *Mycobacterium ulcerans*” *Ecology Letters* (in preparation) Gabriel E. García-Peña, Andrés Garchitorena, **Kevin Carolan**, Benjamin Roche, Jean-François Guégan.
- 3) “Space-time dynamics and landscape-associated risk of Buruli ulcer in Akonolinga, Cameroon.” *PLoS neglected tropical diseases* (2014) Jordi Landier, Jean Gaudart, **Kevin Carolan**, Danny Lo Seen, Jean-François Guegan, Sara Eyangoh, Arnaud Fontanet, Gaëtan Texier.

The first article, Marion *et al.*, contains several maps of the distribution of endemic villages and states in Nigeria and Benin. I geo-referenced these villages, prepared maps of the regions, and assisted in review and comments of the manuscripts.

The second paper, García-Peña *et al.*, explored how *M. ulcerans* is influenced the removal of species from the biotic community, and in general terms the importance of local extinction of species influences multi-host pathogens. In this research, I provided data on deforestation and connectivity between sites, comments and discussion on the manuscript.

The final paper, Landier *et al.* 2014, was a detailed study of the change in the distribution of Buruli ulcer in Akonolinga, through time. I assisted in the generation of data, the land cover maps generated for use in the ecological niche models of *M. ulcerans* were also used here, to explore associations between land cover and Buruli ulcer. In turn I used the prevalence data Jordi collected in the ecological niche of the aquatic insects, to correlate the prevalence of the insects to the distribution of the disease. I also provided reviews and comments of the final manuscript.

## Buruli ulcer in South Western Nigeria

Estelle Marion\* PhD, Kevin Carolan, Ambroise Adeye MD, Marie Kempf PhD, Annick Chauty MD,  
Laurent Marsollier PhD.

Fondation Raoul Follereau, Pobè, Bénin (E. Marion, A. Adeye, A. Chauty)

Inserm, Angers, France (E. Marion, L. Marsollier)

IRD, Montpellier, France (K. Carolan)

Centre Hospitalier Universitaire d'Angers, Angers, France (M. Kempf).

\*Corresponding author, [estelle.marion@inserm.fr](mailto:estelle.marion@inserm.fr), phone: + 33 (2) 44 68 83 14; fax: + 33 (2) 44 68 83

69

**Running title:** Buruli ulcer in South Western Nigeria

**Keywords:** Buruli ulcer, Nigeria, epidemiology

### **Abstract**

Nigeria is known to be endemic to Buruli ulcer, but epidemiological data are remarkably rare. Here, we present a large cohort of Nigerian patients treated in a neighbouring country, Benin. Severe lesions and delay of consultation are factors that should encourage establishment of a treatment centre in South Western Nigeria.

Buruli ulcer, the third most common mycobacterial disease of humans after tuberculosis and leprosy, is an important disfiguring and disabling cutaneous infection caused by *Mycobacterium ulcerans*. Buruli ulcer was declared an emerging skin disease of public health concern by the World Health Organization (WHO) in 1998. Although this mycobacteriosis is known to be linked to swampy areas and environmental changes, the mode of transmission is not yet clearly understood, but the possible role of water bugs has been invoked in the last ten years. Since 2004, the WHO has recommended a daily treatment of rifampin with streptomycin or clarithromycin for 8 weeks as a first line treatment. And surgery is often necessary on large lesions to remove necrotic tissues before carrying out a skin graft. Treatments and wound care need to be done in competent and specialized health structures, which are uncommon in the most Buruli ulcer endemic countries. The disease is widespread in West and Central Africa, French Guiana and much of South America, and Australia, and most of the burden of the disease falls on West and Central Africa, with the majority of cases being in Cote d'Ivoire. In the available epidemiological data, Nigeria presents a notable gap in cases, surrounding countries such as Cameroon and Benin which have highly endemic regions, and this gap is likely due to underreporting and the lack of adequate public health structures.

In Nigeria, Buruli ulcer was first described in 4 patients living in the Benue State in 1967 (Figure 1A). Then 24 Buruli ulcer patients were described in an area around Ibadan town, Oyo State, in 1976 (Figure 1A). 30 years later the Nigerian government, with the support of the WHO, conducted an assessment of Buruli ulcer presence in the Southern and Southeastern states of the country, where cases had been previously reported. 14 patients were considered likely to have a Buruli ulcer, and came from 5 different states: Anambra, Cross river, Enugu, Ebonyi and Akwa Ibom (Figure 1A). Finally, the presence of Buruli ulcer in Nigeria was mentioned in a manuscript, in which 9 *M. ulcerans* strains were isolated from patients living in Oyo, Anambra, Cross river, Enugu, Ebonyi or Ogun states, between year 2006 and 2012 (Figure 1A). In total, 51 Buruli ulcer patients were described during 45 years and were found in Southern Nigeria, which is characterised by a tropical rainforest climate,



similar to Buruli ulcer endemic areas around the Gulf of Guinea. The Northern part of the Nigeria is associated with a tropical dry climate which is not known to be suitable for Buruli ulcer presence.

In this context, we have collected data of Nigerian patients that have been treated in the Buruli ulcer treatment centre of Pobè, Benin, which is located at 3 kilometres from the frontier with Ogun State, Nigeria, between the years of 2004 and 2013.

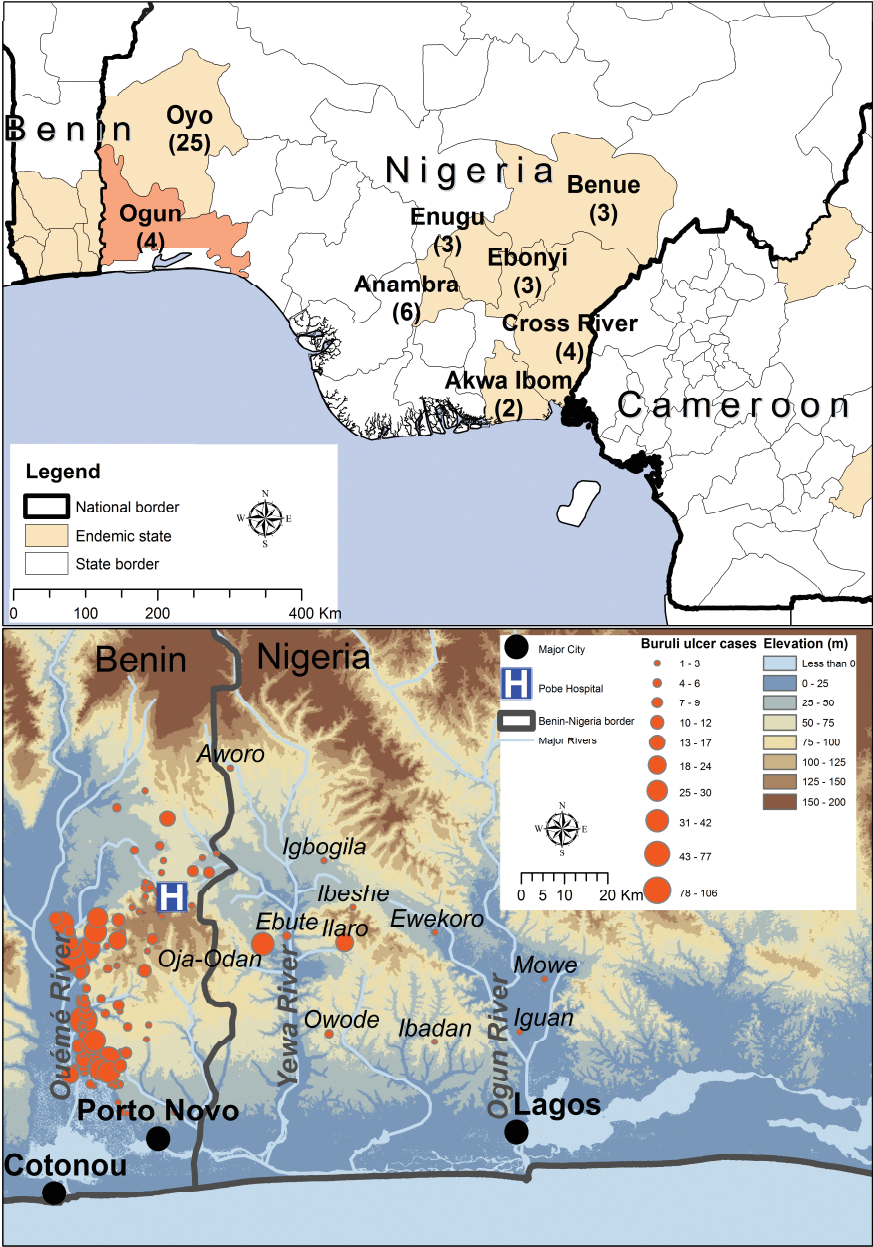


Figure 1. (A) Nigerian districts where Buruli ulcer patients were already described at least once, and neighbouring countries areas where Buruli ulcer is endemic. The number of cases described since 1967 is indicated for each Nigerian district. (B) Location of Benin and Nigerian patients coming in CDTUB-Pobè for treatment of Buruli ulcer.

In 2005 the first Nigerian came to Pobè treatment centre with a Buruli ulcer lesion, one year after the building of the hospital. From 2005 to 2013, 127 Nigerian patients were diagnosed and treated for Buruli ulcer in this hospital and the number of new cases is increasing each year (Figure 2A). There were 52% (66) females and 48% (61) males with a median age of 24 and 10 years, respectively (Figure 2B). Such sex ratio variation with age at diagnosis has been previously reported and could be due to differential exposure to *M. ulcerans*, or different responses by the family and community to infection in boys versus girls. The WHO categorises lesions by size in three categories: category 1, maximum diameter <5 cm, category 2, 5-15cm, and category 3, >15 cm. Nigeria Patients commonly present with large lesions, with 60% (73) of category 3, including 20% with osteomyelitis. There were 8% (10) category 1 and 36% (42) category 2. 78% (99) of lesions were at the ulcerative stage. These data show that Nigerian patients present mainly late stages of Buruli ulcer. This is confirmed by the delay before consultation, where 24% of the patients had waited more than one year between the beginning of the lesion and the consultation in the Buruli ulcer treatment centre in Pobè (Figure 2C). By comparison, among all the Buruli ulcer patients treated for Buruli ulcer between 2005 and 2011 in Pobè treatment center, lesions of category 3 represented only 36% of the lesions.

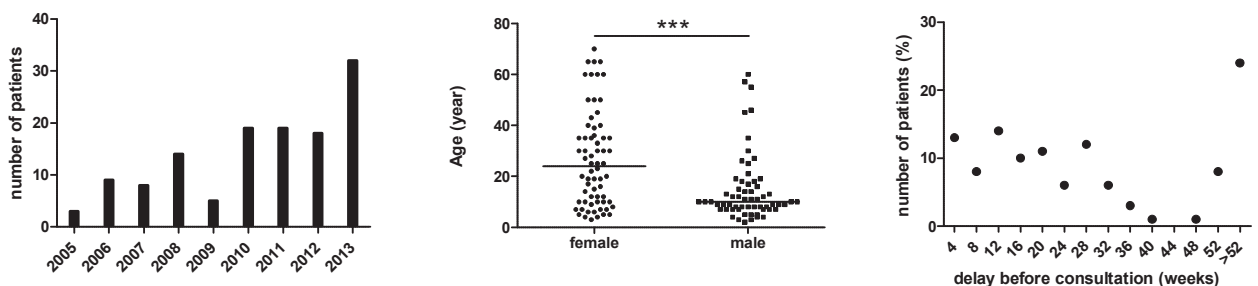


Figure 2. (A) Number of new Buruli ulcer patients coming from Nigeria and treated in Buruli ulcer treatment centre of Pobè, Benin between 2005 and 2013, (B) Age of patients by gender, (C) Distribution of patient consultation delay.

We were able to georeference the origin of 68% (86) of Nigerian patients. They came mainly from the Ogun state which shares its western frontier with Benin (Figure 1B). The furthest village, Mowe, is about 100km from the Buruli ulcer treatment center, and it takes around 6 hours to reach Pobè by motobike. The Nigerian villages reporting the most Buruli ulcer patients are those closest to Pobè: Oja-Odan, Ebute, Ilaro.

The Nigerian endemic area in Ogun state is divided in two drainage basins, the Yewa and the Ogun rivers, while Ouémé /Plateau endemic department in Benin is traversed by the Ouémé river. These 3 draining systems discharge in separate lagoons that are interconnected by channels at the delta. Southeast Benin and Southwest Nigeria do not belong to the same drainage system, but these two Buruli ulcer endemic areas have a similar topography and climate, characterised by tropical rainforest, changeable topography with many small hills and fertile plains. Similar environments are encountered in other Buruli ulcer endemic areas of West Africa, for which the patients are found around different drainage systems but always with broad fertile richly inundates plains.

This report is the first description of a large cohort of PCR-confirmed Buruli ulcer patients coming from South Western Nigeria. As no health centre is dedicated to diagnose and treat Buruli ulcer in this country, patients close to the south western frontier turn to the Beninese health system, where specialised facilities treat Buruli ulcer free of charge or with a small participation. This report supports the establishment of a Buruli ulcer surveillance system in Nigeria, encouraging the training of health workers in Ogun state, and promotes the establishment of a Buruli ulcer treatment centre in South Western Nigeria.

### **Acknowledgments.**

This study was supported by the Fondation Raoul Follereau, the Institut National de la Santé et de la Recherche Médicale (Inserm, Programme Inserm Avenir) and by the Agence Nationale de Recherche sur le SIDA et les Hépatites.

### **Biographical sketch:**

Estelle Marion is a post-doctoral researcher. Her research focuses on the ecology of *M. ulcerans* and host/pathogen interactions. She is also in charge of a molecular biology laboratory in a treatment center of Benin to allow biological confirmation of Buruli ulcer at patient bedside.

Niche-based extinction of species boosts presence of the environmentally-acquired pathogen,  
*Mycobacterium ulcerans*

Gabriel E. García-Peña<sup>1,2,\*</sup>, Andrés Garchitorea<sup>1,3</sup>, Kevin Carolan<sup>1,4</sup>, Elsa Canard<sup>1</sup>, Anne- Hélène Prieur-Richard<sup>5</sup>, Gerardo Suzán<sup>6</sup>, James N. Mills<sup>7,Φ</sup>, Benjamin Roche<sup>8,Φ</sup>, Jean-François Guégan<sup>1,3,Φ</sup>

1. UMR MIVEGEC, Maladies Infectieuses et Vecteurs: Ecologie, Génétique, Evolution et Contrôle. UMR 5290 CNRS-IRD-UM1-UM2, Centre de Recherche IRD, 911 Avenue Agropolis, BP 64501. 34394 Montpellier Cedex 5 – France.
2. Centre de Synthèse et d'Analyse sur la Biodiversité -CESAB. Bâtiment Henri Poincaré, Domaine du Petit Arbois. Avenue Louis Philibert. 13857 Aix-en-Provence Cedex 3 – France.
3. Ecole des Hautes Etudes en Santé Publique, Rennes, France.
4. UMR Territoires, Environnement, Télédétection et Information Spatiale (TETIS) CIRAD, 34093, Montpellier, France
5. DIVERSITAS, Muséum National d'Histoire Naturelle. Paris, France.
6. Departamento de Etología, Fauna Silvestre y Animales de Laboratorio, Facultad de Medicina Veterinaria Zootecnia, Universidad Nacional Autónoma de México, Ciudad Universitaria, México, D.F. C.P. 04510, México.
7. Population Biology, Ecology, and Evolution Program. Emory University, Atlanta, GA, USA.
8. UMMISCO, UMR IRD-UPMC 209, Bondy, France.

Φ JNM, BR and J-FG have equally contributed as Co-PIs of the BIODIS working group.

\* Correspondence: [gabriel.garcia@ird.fr](mailto:gabriel.garcia@ird.fr)

Keywords (10): Community disassemblage, metacommunity, stochastic and niche-based processes, zoonotic disease emergence, diversity loss, multi-host, *Mycobacterium ulcerans*, Buruli ulcer.

Short title (less than 45 characters including spaces): Community disassembly and *Mycobacterium ulcerans*

*Ecology Letters*: maximum of 5000 words in length (main text), 6 figures, tables or boxes, and 50 references.

ABSTRACT (150 words)

Extinction of biota may boost emergence of zoonotic diseases jeopardizing the health of humans and ecosystems. The dilution effect hypothesis suggests that interactions between hosts and non-host species reduce the proportion of hosts infected (prevalence). Hence, extinction of non-hosts and persistence of hosts (niche-based extinction) may increase prevalence. Nevertheless, empirical evidence is elusive because key hosts may go extinct stochastically, and reduce prevalence. Here, a metacommunity approach reveals that niche-based extinction of species increases prevalence of an environmentally-acquired pathogenic bacterium, *Mycobacterium ulcerans*, on its multiple hosts. As a pseudo-experiment, we examined lentic communities assembled by niche-based processes and lotic communities assembled by stochastic processes. Accounting for stochastic processes, prevalence of the bacilli increased when species were locally extinct in both, lentic and lotic, metacommunities. Thus empirical evidence supported the dilution effect hypothesis and highlighted the importance of researching emerging infections with a metacommunity perspective that considers niche-based and stochastic processes.

## INTRODUCTION

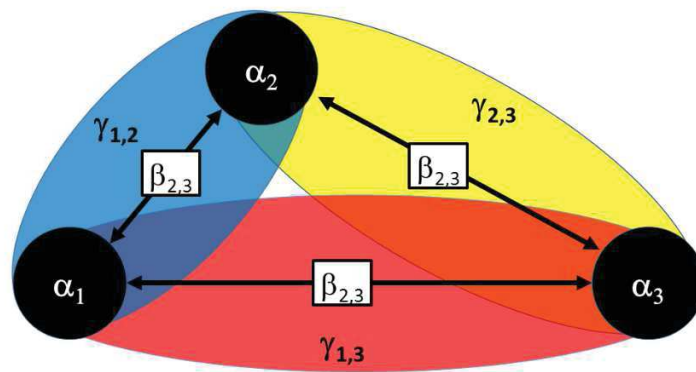
Currently, a sixth mass extinction is occurring in the planet and species are declining locally at a high rate (Dirzo *et al.* 2014). Amid such events, preserving ecosystem health and human welfare requires understanding the consequences of local extinction of species on ecosystem function (Rafaelli 2004; Balvanera *et al.* 2006). Consequences of concern are the effects that eroding biodiversity produces on the emergence of zoonotic diseases in humans and wildlife (Harvell *et al.* 2002; Ostfeld *et al.* 2008). In this regard, the dilution effect hypothesis suggests that the percentage of individuals infected by a pathogen –henceforth prevalence– is influenced by inter-specific interactions that hinder infection of susceptible hosts and increase recovery of infected hosts (Keesing *et al.* 2006; Ostfeld *et al.* 2008). Hence, species diversity may reduce prevalence, and prevalence may boost if these inter-specific interactions are impaired by the local extinction of hosts and non-host species (Ezenwa *et al.* 2006; Suzán *et al.* 2009). Nevertheless, the underlying mechanisms of the dilution effect are controversial and elusive (Leibold *et al.* 2004; Randolph & Dobson 2012; Ostfeld 2013; Roche *et al.* 2013b). Further evidence suggests that the effects of biodiversity loss on prevalence are sensitive to disassembly rules determining the order at which host species go extinct in the community (Ostfeld & LoGiudice 2003; Lafferty 2012; Lacroix *et al.* 2013). For example, Ostfeld and LoGiudice (2003) found two contrasting relationships in a theoretical study. On the one hand, random extinction of host species reduced prevalence of nymphal ticks infected with *Borrelia burgdorferi*. On the other hand, extinction of species increased prevalence substantially when applying the disassembly rule that competent hosts (mice) were the last species to go extinct. These contrasting findings rest on the assumptions that local extinction of species is niche-based and competent hosts are persistent within local communities (Randolph & Dobson 2012; Johnson *et al.* 2013; Joseph *et al.* 2013). Therefore the generalization of these hypotheses must be investigated in the light of empirical evidence and integrative methods of community ecology (Leibold *et al.* 2004; Chase & Myers 2011).

Communities are assembled and disassembled by both, (1) stochastic processes such as random extinction of species and probabilistic (passive) dispersal; and (2) niche-based processes including symbiotic interactions and species tolerance to abiotic conditions (Tokeshi 1999; Leibold *et al.* 2004; Helmus *et al.* 2007; Chase & Myers 2011; Moritz *et al.* 2013). In such instance, predicting prevalence is daunting because host species may colonize and go locally extinct by the contribution of both, niche-based and stochastic processes, which are contingent on interconnectivity between local communities, ecosystem types, taxonomic groups, life history strategies, and inter-specific interactions (Diamond 1975; Connor & Simberloff 1979; Ricklefs 1987; Hubbell 2001; Bennett & Owens 2002). However, these processes can be analysed at the

multiscale perspective of the metacommunity (Leibold *et al.* 2004), which is rarely used in epidemiological studies (Mihaljevic 2012). In principle, local ( $\alpha$ ) diversity of species in a community would be determined by regional ( $\gamma$ ) diversity in the metacommunity, and processes affecting  $\beta$ -diversity, i.e. the dissimilarity between communities (Whittaker 1972; Loreau 2000; Leibold *et al.* 2004; Chase & Myers 2011).  $\beta$ -diversity can be decomposed into two distinct informative components: species replacement (turnover) and loss of species in communities. Loss of species in communities tends to generate nestedness of communities so that the poorest community becomes a subset of the most diverse community (Ulrich & Almeida-Neto 2012; Baselga 2013). Furthermore, loss of species may be both random (stochastic) and niche-based, the latter being ordered by species tolerance to abiotic conditions and symbiotic interactions (Tokeshi 1999; Leibold *et al.* 2004; Helmus *et al.* 2007; Chase & Myers 2011; Moritz *et al.* 2013). Thus, if prevalence is boosted due to niche-based extinction, prevalence should positively correlate to the extent of nestedness once that stochastic effects are accounted for.

We tested the above prediction by using a metacommunity framework (Fig. 1) to examine aquatic communities of tropical Africa and prevalence of the pathogenic bacterium *Mycobacterium ulcerans* (MU). MU produces a cutaneous necrotizing infection (Buruli ulcer) in humans inhabiting tropical rural areas in proximity to slow-flowing watercourses and stagnant water, and experiencing rapid environmental changes due to deforestation, agriculture and aquaculture (Wansbrough-Jones & Phillips 2006; Brou *et al.* 2008; Walsh *et al.* 2008; Merritt *et al.* 2010). Presence of MU is dependent on the composition of aquatic invertebrates and functional groups in the community (Benbow *et al.* 2013). However, it is unknown if these host species can amplify and transmit MU, or if they simply act as host carriers -henceforth hosts (Roche *et al.* 2013a). Here we investigated presence of MU on the five most represented and persistent orders of macro-invertebrates (Coleoptera, Diptera, Ephemeroptera, Hemiptera and Odonata), some of which are considered amplifiers of MU, e.g. Hemiptera (Portaels *et al.* 2001; Marsollier *et al.* 2003). To investigate diversity in the aquatic communities, we used a comprehensive and unique database for a Cameroon area near Akonolinga (Garchitorena *et al.* 2014). Within this region, two types of communities were recognized: seven communities with stagnant or very slow flowing water ponds (lentic metacommunity), and nine communities mainly with water currents of the Nyong and Mfoumou rivers (lotic metacommunity). A previous study suggested that the transformation of lentic systems into lotic systems, during 27 years, encompassed a shift in the processes assembling communities, from niche-based processes towards stochastic ones (Brown & Milner 2012). These observations were corroborated by metacommunity analyses on our data (Supplementary materials

SM1). Niche-based processes dominated the assembly of lentic communities whereas stochastic processes dominated the assembly of lotic communities. Hence, we analysed the lotic and lentic communities as two separated metacommunities, and examined the results as a pseudo-experiment to assess the effects of stochastic processes on the relationship between nestedness and MU prevalence. In general, we found two main results. First, opposite to the prediction of the dilution effect, generalized linear models (GLMs) revealed that MU prevalence was positively correlated with  $\alpha$ -diversity and  $\gamma$ -diversity in the lotic metacommunity, but not in the lentic one. This suggested that stochastic processes may exert important effects on the relationship between prevalence and diversity. Finally, as expected, we found that MU prevalence at regional level was positively correlated with nestedness in both, the lentic metacommunity and in the lotic metacommunity, once that we accounted for stochastic processes represented by probabilistic dispersal of species between communities.



**Figure 1.** Metacommunity framework used to investigate regional prevalence of *Mycobacterium ulcerans*. In this example, each of three regions (blue, red, and yellow) encompasses two communities ( $i$  and  $j$ ) with their respective local diversities ( $\alpha_i$  and  $\alpha_j$ ). Regional prevalence $_{i,j}$  is influenced by regional diversity ( $\gamma_{i,j}$ ), and stochastic and niche-based processes driving dissimilarity between the two communities ( $\beta_{i,j}$ ).  $\beta_{i,j}$  is decomposed in two informative components: ordered loss of species (nestedness) and turnover of species between communities. Both, niche-based and stochastic extinctions tends to increase the extent of nestedness between communities. Thus, we examine nestedness and regional MU prevalence $_{i,j}$  while accounting for stochastic processes assembling communities: probabilistic dispersal determined by the connectivity between communities (arrows) and the total abundance of animals in the region, regardless the species.



## METHODS

MU is a relatively slow growing bacterium (~ 50h for replication) that synthesizes a lipophilic macrolide toxin (mycolactone), and evolved multiple DNA deletions and rearrangements associated with niche reduction (Stinear *et al.* 2007). Some works support the existence of vector-borne transmission and amplification by Belostomatidae and Naucoridae water bugs (Marsollier *et al.* 2003), and previous studies suggest that MU is embedded within ecological networks (Roche *et al.* 2013a). MU occurs on a wide diversity of substrata such as mud, organic detritus, biofilms on aquatic plants and in a large diversity of aquatic micro- and macro-invertebrates, fish, amphibians, reptiles and mammals (Portaels *et al.* 2001; Marsollier *et al.* 2004; Benbow *et al.* 2013; Willson *et al.* 2013; Garchitorena *et al.* 2014; Morris *et al.* 2014).

### *Sampling and study sites*

Periodic sampling of aquatic communities was performed in Akonolinga, Cameroon between June 2012 and May 2013. Monthly samples were collected in 16 aquatic communities within a region of approximately 3,600 km<sup>2</sup> including a wide spectrum of streams, rivers, swamps and flooded areas. Sampling was performed during five consecutive days, and in each aquatic community four locations were chosen in areas of slow water flow and dominant aquatic vegetation. Classification of aquatic invertebrates was performed at the family level (see Garchitorena *et al.* 2014 for details).

### *Detection of *M. ulcerans* with qPCR*

For each community and month sampled, the presence of MU was tested on 17 different taxonomic groups of animals (Garchitorena *et al.* 2014). However, not all of these groups were present across the study, and this may add variation in the presence of MU due to the differences across taxa in the affinity to harbour the bacterium (Portaels *et al.* 2001; Marsollier *et al.* 2004). To assure consistency in the sensitivity of detecting MU we focused on the orders of macro-invertebrates that were consistently present during the sampling year (Coleoptera, Diptera, Ephemeroptera, Hemiptera and Odonata, SM2). Quantitative PCR was used to detect two specific markers of MU in each sample: oligonucleotide primer and TaqMan probe sequences of IS2404 and the ketoreductase B domain of the mycolactone polyketide synthase (mls) gene from the plasmid pMUM001 (Garchitorena *et al.* 2014).

### *Probabilistic dispersal*

Theory predicts that stochastic extinction of species in a community should decrease prevalence, as opposed to niche-based extinction of species that are expected to increase pathogen transmission (Ostfeld & LoGiudice 2003). Stochastic extinctions and colonisations of species in communities may occur due to probabilistic dispersal of species between neighbour communities (Connor & Simberloff 1979; Hubbell 2001; Chase & Myers 2011). In such instance, nearest communities are expected to be more similar than distant ones. Thus, we considered the connectivity between communities as surrogate of probabilistic dispersal, and accounted for its effects in our analyses in three ways: First we assessed whether our data (local MU prevalence and  $\alpha$ -diversity) were structured by the connectivity between communities (SM3). Second, we included estimates of connectivity as predictors in the GLMs on regional MU prevalence (see below). Finally, because connectivity may interact synergistically with the other predictors of interest (i.e. nestedness and  $\gamma$ -diversity) we tested for two-way interactions between connectivity and the other predictors included in the GLMs on regional MU prevalence.

Connectivity between communities was represented by two estimates: Euclidean connectivity and constrained connectivity (SM3). Euclidean connectivity assumed that host and non-host species can disperse freely between communities across the landscape. Constrained connectivity assumed that species disperse within the hydrological system in Akonolinga. Connectivity was measured as the inverse of distances between communities (Euclidean and constrained):  $\sqrt{1 - \frac{d}{d_{max}}}$ , from (0) no connectivity to (1) full connectivity. We present separate GLMs for each estimate of connectivity because including both estimates as predictors in the same GLM produced multicollinearity in the models.

### *Local analyses on MU prevalence*

We performed GLMs to test the correlation between MU prevalence (the proportion of positive samples found during a year in each community) and  $\alpha$ -diversity represented by the Shannon index ( $H$ ). The Shannon index was estimated as:  $H = 1 - \sum_{i=1}^S p_i \ln(p_i)$ ; where  $p_i$  was the proportional abundance of host species  $i$  and  $S$  was the total number of families in the community (Shannon 1948). Additionally, we considered two potential confounding effects that may influence MU prevalence: deforestation and total abundance of animals in the community, regardless the taxonomic group -henceforth abundance. Deforestation is known to affect Buruli ulcer (Landier *et*

al. 2014) and may thus influence MU prevalence as well. To account for this confounding effect in the analyses we considered deforestation as the area deforested within a circular perimeter of 1km radius around each community, between the years 2000-2011 (details in SM4). Abundance was used as surrogate of productivity and/or the accumulation of animals due to physical forces unrelated to ecological/taxonomic diversity. GLMs assumed a binomial distribution of the response variable, predictors were standardized ( $\ln x - \ln x_{\text{mean}}$ ), and normality of residuals was assessed by graphical exploration and Shapiro-Wilk' tests (SM5). Because multicollinearity may be due to correlation between predictor variables, we ensured that the variance inflation factors (VIF) of each parameter included in the models were  $< 5$  (Heiberger & Holland 2004). Abundance and Shannon index produced multicollinearity, hence we performed separate GLMs to analyse their correlations with the MU prevalence. Finally, we ensured that the quality of our results was not sensitive to either the taxonomic variation of the samples in which MU was detected (SM2.1-2.2), and spatial autocorrelation (SM3).

#### *Regional analyses on MU prevalence*

We used stepwise model selection of GLMs based on the AIC, Akaike information criterion (Burnham & Anderson 2002), to test correlations between regional MU prevalence and  $\gamma$ -diversity, nestedness, connectivity and confounding factors: turnover, deforestation and total abundance of animals, henceforth abundance at regional level. Regional MU prevalence,  $\gamma$ -diversity ( $H$ ), abundance and deforestation were estimated by pooling data of each pair of communities, for all possible pairwise combinations of lentic communities and all possible pairwise combinations of lotic communities. Because  $\gamma$ -diversity, abundance and deforestation are magnitudes in different scales, they were standardized as in the local analyses. Surrogates of nestedness and turnover between paired communities were estimated from the decomposition of the Bray-Curtis dissimilarity proposed by Baselga (2013):  $\beta_{\text{balanced}}$  was associated with turnover, capturing abundance of species being replaced by abundance of other species; and  $\beta_{\text{gradient}}$  was associated with nestedness of communities, capturing unidirectional changes in the abundance of species (Fig. 2, SM6). We included turnover ( $\beta_{\text{balanced}}$ ) as a confounding effect in the GLMs because MU prevalence may be sensitive to the taxonomic composition of the communities (Benbow *et al.* 2013; Roche *et al.* 2013a). While performing model selection we tested for correlations between MU prevalence and the two-way interactions between connectivity (Euclidean or constrained) and all the other predictors in the GLMs. Multicollinearity was avoided by assuring VIFs  $< 5$ , and normal distribution of residuals was assessed (SM5). The VIF criteria excluded models of lotic communities including both predictors:  $\gamma$ -diversity and the regional abundance of animals. Thus,

we tested each predictor in separate GLMs. Like in the analyses on local prevalence, we corroborated that our results were not sensitive to the taxonomic variation of the sample, by adding the taxonomic order of the host as random factor in the GLMs (SM2). Finally, we ensured that the results were not biased due to the contribution of a particular community by removing one community from the data and repeating the analysis (SM7-8). This procedure was repeated for each of the communities. All analyses were performed in R software (R-Development-Core-Team 2011). For all GLMs we present estimates (b)  $\pm$  standard errors, and a Z test with the probability of accepting the null hypotheses that  $b = 0$ .

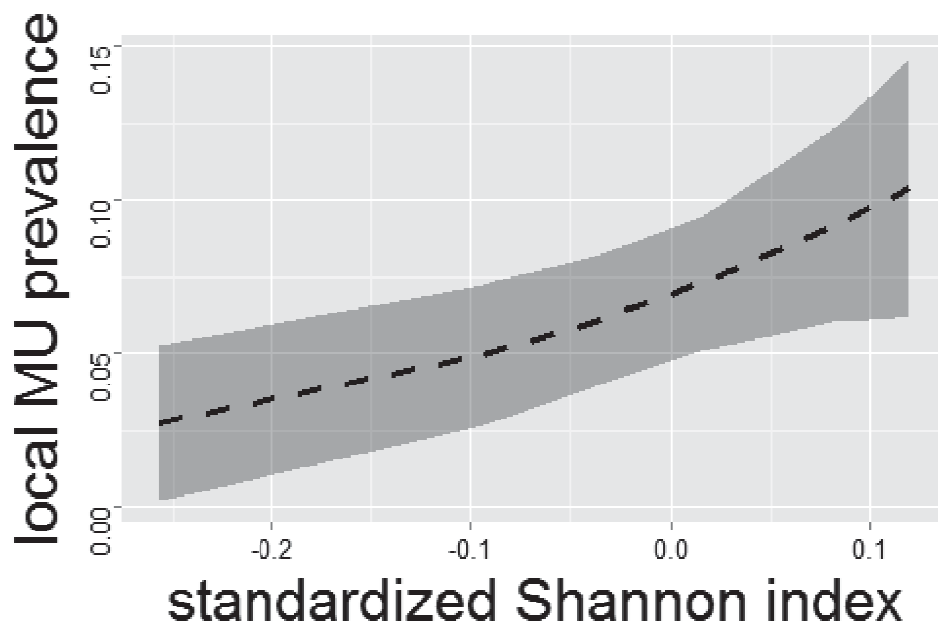
		sp1	sp2	sp3	sp4
A)	Site 1	20	20	10	10
	Site 2	10	10	20	20
B)	Site 1	20	20	20	20
	Site 2	10	10	10	10
C)	Site 1	20	20	20	20
	Site 2	20	20		

**Figure 2.** Decomposition of Bray-Curtis dissimilarity ( $\beta$ ). In all instances  $\beta = 0.33$ . **A** represent turnover of species abundance ( $\beta_{\text{balanced}} = 0.33$ , and  $\beta_{\text{gradient}} = 0$ ); **B** represents unidirectional changes in abundance ( $\beta_{\text{balanced}} = 0$ , and  $\beta_{\text{gradient}} = 0.33$ ); and **C** represents nestedness in which site 2 is a subset of site 1 ( $\beta_{\text{balanced}} = 0$ , and  $\beta_{\text{gradient}} = 0.33$ ). Modified from Baselga (2013).

## RESULTS

### *Local analyses on MU prevalence*

At the local level, MU prevalence in the lentic communities was not significantly correlated with  $\alpha$ -diversity ( $b = -0.81 \pm 1.44$ ,  $Z = -0.6$ ,  $p = 0.575$ ) and deforestation ( $b = -0.19 \pm 0.32$ ,  $Z = -0.63$ ,  $p = 0.529$ ). However, MU prevalence in the lotic communities was positively correlated with  $\alpha$ -diversity ( $b = 3.84 \pm 1.65$ ,  $Z = 2.33$ ,  $p = 0.020$ , Fig. 3) but not with deforestation ( $b = -0.08 \pm 0.28$ ,  $Z = -0.30$ ,  $p = 0.768$ ). The relationship between total abundance of animals and MU prevalence was analysed in separated GLMs, and we found a negative relationship in the lotic communities ( $b = -0.84 \pm 0.37$ ,  $Z = -2.29$ ,  $p = 0.022$ ), and no significant relationship in the lentic communities ( $b = -0.22 \pm 0.30$ ,  $Z = -0.72$ ,  $p = 0.469$ ).



**Figure 3.** Positive relationship between local prevalence of *M. ulcerans* (predicted from GLMs) and the Shannon index of  $\alpha$ -diversity in lotic communities: mean values (solid line) and upper and lower 95% confidence intervals (grey shaded area) are represented.

### Regional analyses on MU prevalence

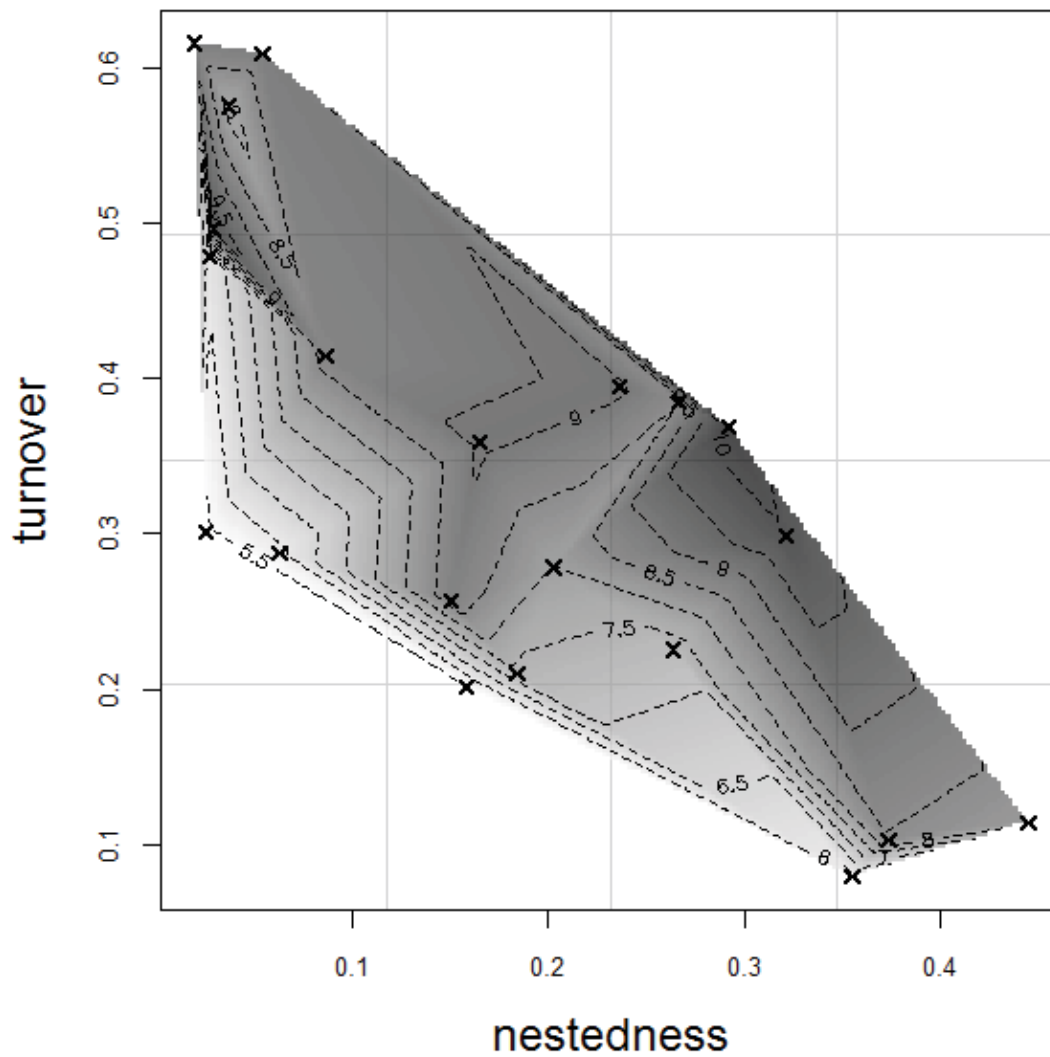
Stepwise model selection suggested that regional MU prevalence was positively correlated with  $\gamma$ -diversity in the lotic metacommunity whereas in the lentic metacommunity the correlation was not significant (Tables 1, 2). Regional MU prevalence was positively correlated with nestedness in the lentic metacommunity (Table 1, Fig. 4). However, in the lotic metacommunity this correlation was either significant or not significant (Table 2A-D). The significance of this correlation relies upon the effects of connectivity between communities, and the surrogate of connectivity used as predictor in the models. While MU prevalence was not significantly correlated with the interaction between nestedness and constrained connectivity (Table 2A), it was positively correlated with the interaction between nestedness and Euclidean connectivity (Table 2B, 2D). This correlation supported our prediction suggesting that an increase in regional MU prevalence was associated with an increase in nestedness. Additionally, we found remarkable confounding effects in our GLMs. In both lentic and lotic metacommunities, an increase in regional abundance of animals was associated with a decrease in regional MU prevalence. In the lentic metacommunity, increments in regional deforestation and turnover were associated with an increase in MU prevalence (Table 1, Fig. 4); and in the lotic metacommunity MU prevalence decreased when turnover increased (Table 2B).

**Table 1.** Generalized linear models fitted on regional prevalence of *Mycobacterium ulcerans* (response variable) in 21 pairs of lentic communities. The final model (AIC = 130.5) was selected by stepwise (backward and forward) selection starting from the model with all principal effects (AIC = 131.8). All possible interactions between connectivity (constrained and Euclidean) and the other predictors were tested. Estimates (b), standard errors, Z tests supporting that  $b \neq 0$ , and (VIF) variance inflation factor are presented. Results did not change in quality when using AICc penalization for small sample sizes during stepwise model selection. Significant parameters are presented in bold. Legend:  $\gamma$ -diversity is the Shannon index calculated for each pair of lentic communities, abundance is the total abundance of animals in each sample, regardless the species, turnover ( $\beta_{\text{balanced}}$ ) is the replacement of species between pair of communities, and nestedness ( $\beta_{\text{gradient}}$ ) between each pair of communities.

	starting model			final model	
predictors	<b>b ± SE</b>	<b>Z (p)</b>	<b>VIF</b>	<b>b ± SE</b>	<b>Z (p)</b>
$\gamma$ -diversity	-0.85 ± 0.74	-1.15 (0.252)	1.21		
abundance	<b>-0.76 ± 0.28</b>	<b>-2.70 (0.007)</b>	<b>1.59</b>	<b>-0.57 ± 0.25</b>	<b>-2.29 (0.022)</b>
deforest	<b>0.29 ± 0.14</b>	<b>2.06 (0.040)</b>	<b>1.60</b>	<b>0.25 ± 0.13</b>	<b>1.98 (0.047)</b>
connectivity <sub>constrained</sub>	0.50 ± 0.47	1.06 (0.291)	1.56		
turnover	<b>1.68 ± 0.68</b>	<b>2.48 (0.013)</b>	<b>2.94</b>	<b>1.58 ± 0.60</b>	<b>2.66 (0.008)</b>
nestedness	<b>1.19 ± 0.73</b>	<b>1.62 (0.105)</b>	<b>2.51</b>	<b>1.42 ± 0.69</b>	<b>2.04 (0.041)</b>

**Table 2.** Generalized linear models fitted on regional prevalence of *M. ulcerans* (response variable) in 36 pairs of lotic communities. Based on AIC, the final models were selected by stepwise selection of variables from the starting models containing only the principal terms (A and B models included  $\gamma$ -diversity, and C and D models included abundance as predictors in the starting model; see SM 9). All two-way interactions between connectivity (constrained and Euclidean) and the other predictors were tested. All results presented here did not change in quality when using AICc penalization for small sample sizes during stepwise model selection. Abundance and  $\gamma$ -diversity were not included in the same GLM because they produced multicollinearity. Likewise, constrained connectivity and Euclidean connectivity were analysed in separate GLMs. Significant parameters are presented in bold. Estimates b, standard errors, and Z tests supporting that  $b \neq 0$ , are presented. Legend: see Table 1.

Selected model (AIC)	A (174.1)		B (173.5)		C (176.8)		D (177.1)	
Predictors	<b>b</b> $\pm$ SE	<b>Z</b> (p)	<b>b</b> $\pm$ SE	<b>Z</b> (p)	<b>b</b> $\pm$ SE	<b>Z</b> (p)	<b>b</b> $\pm$ SE	<b>Z</b> (p)
$\gamma$ -diversity	<b>3.06 + 0.68</b>	<b>4.51 (&lt; 0.001)</b>	<b>2.53 + 0.85</b>	<b>2.99 (0.003)</b>				
abundance					<b>-0.46 + 0.13</b>	<b>-3.69 (&lt; 0.001)</b>	<b>-0.45 + 0.15</b>	<b>-3.01 (0.003)</b>
deforestation					-0.20 + 0.13	-1.53 (0.125)	-0.24 + 0.14	-1.78 (0.076)
connectivity constrained (c)	<b>1.47 + 0.73</b>	<b>2.01 (0.044)</b>						
connectivity euclidean (e)			-0.64 + 0.44	-1.45 (0.146)			-0.25 + 0.41	-0.61 (0.544)
turnover (t)	1.53 + 1.07	1.43 (0.152)	<b>-1.14 + 0.48</b>	<b>-2.36 (0.018)</b>				
nestedness (n)			<b>-4.68 + 1.87</b>	<b>-2.51 (0.012)</b>			<b>-3.04 + 1.51</b>	<b>-2.02 (0.044)</b>
c:t	<b>-3.31 + 1.68</b>	<b>-1.97 (0.048)</b>						
e:n			<b>5.28 + 2.24</b>	<b>2.36 (0.018)</b>			3.51 + 1.99	1.76 (0.078)



**Figure 4.** Relationships between regional prevalence of *M. ulcerans* and turnover and nestedness ( $\beta_{\text{gradient}}$ ) in the lentic metacocommunity. The percentage of individuals infected by MU, predicted by GLMs, is presented as a gradient from light gray (lowest probability) to dark gray (highest probability). The observed values are presented by a cross and dashed lines show isolines of prevalence values. In general, regional MU prevalence increases in correlation with both turnover and nestedness.



## DISCUSSION

Our results contradict the negative relationship between prevalence and diversity that the dilution effect hypothesis predicts (Keesing *et al.* 2006). Analyses at a local and regional scales reveal positive relationships between MU prevalence and  $\alpha$ -diversity and  $\gamma$ -diversity in lotic communities of tropical Africa, which are consistent with the finding of a previous study in Ghana at a larger scale (Benbow *et al.* 2013). Controversial results supporting or rejecting the dilution effect are not new in the scientific literature (Randolph & Dobson 2012; Ostfeld 2013; Roche *et al.* 2013b). While the dilution effect hypothesis suggests that inter-specific interactions reduce prevalence of multi-host pathogens (Keesing *et al.* 2006), alternative hypotheses suggests that in some circumstances these interactions may boost prevalence (Randolph & Dobson 2012; Roche *et al.* 2013b). On this controversy, our study adds evidence that the contribution of stochastic processes of community assembly, unrelated to inter-specific interactions, may also be key to understand the relationships between diversity and prevalence. Stochastic processes of community assemblage are more important in determining diversity in the lotic metacommunity than in the lentic metacommunity (SM1, Brown & Milner 2012); and correspondingly the positive relationship between diversity and MU prevalence is observed for the lotic metacommunity but not for the lentic one. Thus the possibility that MU prevalence and species diversity increase together in lotic communities, without encompassing key inter-specific interactions, should not be disregarded. The same stochastic processes that drive diversity may influence MU bacteria, hence increase of multi-host bacteria could be expected when diversity increases. Furthermore, two results of our regional analyses provide evidence supporting that stochastic processes may affect MU prevalence (Tables 1, 2). First, in both communities, lentic and lotic, regional MU prevalence decreases when the regional abundance of animals increases. These negative relationships contradict our expectations, and their underlying mechanisms deserve further research. Possible explanations may be the effects of water currents removing invertebrates, notably larvae or adults living at the bottom or in the middle of the water column, out of the communities and detaching MU from detritus; and/or that the slow growth rate of MU entails a slow infection rate, such that increasing abundance of animals reduces the probability of finding infected animals. Second, the surrogates of probabilistic dispersal (connectivity between communities) have substantial effects on regional MU prevalence in the lotic but not in the lentic metacommunities. In the lotic metacommunity, these surrogates interact with our estimates of  $\beta$ -diversity (turnover and nestedness) explaining substantial variation in regional MU prevalence (Table 2).

Our metacommunity approach (Figure 1) allows us to investigate evidence that niche-based processes influence MU prevalence. First, we find support to our hypothesis because regional MU prevalence positively correlates to nestedness between communities (Tables 1, 2). Remarkably, in the lotic metacommunity this relationship is influenced by the interaction between nestedness and probabilistic dispersal, the Euclidean connectivity between communities (Table 2B, 2D). According to Ostfeld and LoGiudice (2003), niche-based extinction with persistence of host species in communities produces a boost in prevalence of the pathogen causing Lyme disease. In our study we cannot ensure that the hosts are persistent because it is unknown which host species can amplify and transmit MU and/or act simply as host carriers (Roche *et al.* 2013a). Nevertheless, MU was detected in the order Hemiptera and four other orders that were persistent across communities in our study. Previous studies suggest the amplification of MU favoured by Hemiptera water bugs: Belostomatidae and Naucoridae (Marsollier *et al.* 2003). Hence, it may be possible that indeed the hosts of MU are persistent across our study. Finally, regional MU prevalence in lentic communities is correlated with niche-based processes represented by species turnover (Tables 1-2A). Currently, the niche of MU is still poorly understood, making it difficult to disentangle how replacement of host species may affect regional prevalence. While replacing species may involve changes in ecosystem function, it is not clear why turnover may boost or reduce MU prevalence. One plausible explanation is that some key species may act as amplifiers (e.g., Naucoridae and Belostomatidae water bugs) and boost prevalence while others may act as dilutors (e.g., Gastropods) and reduce prevalence (Marsollier *et al.* 2003; Marsollier *et al.* 2004). Likewise, other studies reveal that MU prevalence is associated with host species composition (Benbow *et al.* 2013), and the occurrence of some functional groups, such as filters and scavengers (Roche *et al.* 2013a). Our results provide evidence supporting these insightful studies, by suggesting that different taxa have different effects on MU prevalence. Indeed, investigating the symbiotic relationships between host species in the community and the bacteria will be a fundamental step towards mitigating Buruli ulcer. In this regard, intervention against deforestation may be key to reduce MU prevalence, as we find a positive relationship between these two variables in the lentic metacommunity (Table 1).

## CONCLUSION

Taken together, our study provides firm evidence on the importance of differentiating niche-based processes and stochastic processes of community assembly to understand how local extinction of species may influence circulating multi-host pathogens. By using a metacommunity perspective we find empirical evidence supporting the theoretical study of Ostfeld and LoGuidice (2003). Niche-based extinction of species boosts prevalence of *Mycobacterium ulcerans*, however, the stochastic processes influencing diversity must be accounted for because they also influence the relationships between diversity and microbial prevalence. Furthermore, our study supports the insight that microbial prevalence is influenced by community composition, rather than biodiversity *per se* (Randolph & Dobson 2012; Benbow *et al.* 2013; Roche *et al.* 2013b). Hence, results suggest that direct or indirect impacts exerted on host diversity and habitats might have strong influence on the dynamics of environmentally-persistent microbial agents, and that understanding and forecasting emerging infectious diseases requires fundamental research on niche-based and stochastic processes that influence host communities and microbial forms within a whole metacommunity.

## ACKNOWLEDGEMENTS

This work is supported by the Fondation pour la Recherche sur la Biodiversité (FRB) through its Centre de Synthèse et d'Analyse sur la Biodiversité (CESAB) and its funded working group programme “Disentangling the linkages between biodiversity and emerging infectious diseases” (BIODIS), and with additional support from the international research programme DIVERSITAS. GEGP received a post-doctoral fellowship from FRB/CESAB. GEGP and GS are members of Asociación Mexicana de Medicina de la Conservación (Kalaankab). BR, KC, EC and J-FG were funded by the Laboratoire d'Excellence Centre d'Etude de la Biodiversité Amazonienne (grant ANR-10-LABX-25-01). AG, KC, BR and JFG were supported by a grant from the French National Research Agency (ANR 11 CEPL 007 04 EXTRA-MU). We also thank the BIODIS working-group from CESAB/FRB for insightful discussions on this study.

## REFERENCES

- Balvanera P., Pfisterer B.A. & Buchmann N. (2006). Quantifying the evidence for biodiversity effects on ecosystem functioning and services. *Ecology Letters*, 9, 1146-1156.
- Baselga A. (2013). Separating the two components of abundance-based dissimilarity: balanced changes in abundance vs. abundance gradients. *Methods in Ecology and Evolution*, 4, 552-557.
- Benbow M.E., Kimbirauskas R., McIntosh M.D., Williamson H., Quaye C., Boakye D., Small P.L.C. & Merritt R.W. (2013). Aquatic Macroinvertebrate Assemblages of Ghana, West Africa: Understanding the Ecology of a Neglected Tropical Disease. *EcoHealth*.
- Bennett P.M. & Owens I.P.F. (2002). *Evolutionary Ecology of Birds: Life Histories, Mating Systems, and Extinction*. Oxford University Press, Oxford.
- Brou T., Broutin H., Elguero E., Asse H. & Guegan J.-F. (2008). Landscape Diversity Related to Buruli Ulcer Disease in Côte d'Ivoire. *PLoS Neglected Tropical Diseases*, 2, e271.
- Brown L.E. & Milner A.M. (2012). Rapid loss of glacial ice reveals community assembly processes. *Global Change Biology*, 18, 2195-2204.
- Burnham K.P. & Anderson D.R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2 edn. Springer-Verlag, New York, USA.
- Connor E.F. & Simberloff D. (1979). The assembly of species communities: chance or competition? *Ecology*, 60, 1132-1140.
- Chase J.M. & Myers J.A. (2011). Disentangling the importance of ecological niches from stochastic processes across scales. *Philosophical Transactions of the Royal Society B*, 366, 2351-2363.
- Diamond J.M. (1975). Assembly of species communities. In: *Ecology and Evolution of Communities* (eds. Cody ML & Diamond JM). Belknap Press, Harvard University Press Cambridge, MA, pp. 342-444.
- Dirzo R., Young H.S., Galetti M., Ceballos G., Isaac N.J.B. & Collen B. (2014). Defaunation in the Anthropocene. *Science*, 345, 401-406.
- Ezenwa V.O., Godsey M.S., King R.J. & Guptill S.C. (2006). Avian diversity and West Nile virus: testing associations between biodiversity and infectious disease risk. *Proceedings of The Royal Society Biological Sciences*, 273, 109-117.
- Garchitorena A., Roche B., Kamgang R., Ossomba J., Babonneau J., Landier J., Fontanet A., Flahault A., Eyangoh S., Guégan J.-F. & Marsollier L. (2014). Mycobacterium ulcerans Ecological Dynamics and Its Association with Freshwater Ecosystems and Aquatic Communities:

- Results from a 12-Month Environmental Survey in Cameroon. *PLoS Neglected Tropical Diseases*, 8, e2879.
- Harvell C.D., Mitchell C.E., Ward J.R., Altizer S., Dobson A.P., Ostfeld R.S. & Samuel M.D. (2002). Climate warming and disease risks for terrestrial and marine biota. *Science*, 296, 2158-2162.
- Heiberger R.M. & Holland B. (2004). *Statistical Analysis and Data Display: An Intermediate Course with Examples in S-Plus, R, and SAS*. Springer, New York, USA.
- Helmus M.R., Savage K., Diebel M.W., Maxted J.T. & Ives A.R. (2007). Separating the determinants of phylogenetic community structure. *Ecology Letters*, 10, 917-925.
- Hubbell S.P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, NJ, USA.
- Johnson P.T.J., Preston D.L., Hoverman J.T. & Richgels K.L.D. (2013). Biodiversity decreases disease through predictable changes in host community competence. *Nature*.
- Joseph M.B., Mihaljevic J.R., Orlofske S.A. & Paull S.H. (2013). Does life history mediate changing disease risk when communities disassemble? *Ecology Letters*, 16, 1405-1412.
- Keesing F., Holt R.D. & Ostfeld R.S. (2006). Effects of species diversity on disease risk. *Ecology Letters*, 9, 485-498.
- Lacroix C., Jolles A., Seabloom E.W., Power A.G., Mitchell C.E. & Borer E.T. (2013). Non-random biodiversity loss underlies predictable increases in viral disease prevalence. *Journal of The Royal Society Interface*, 11, 20130947.
- Lafferty K.D. (2012). Biodiversity loss decreases parasite diversity: theory and patterns. *Philosophical Transactions of the Royal Society B*, 367, 2814-2827.
- Landier J., Gaudart J., Carolan K., Lo-Seen D., Guégan J.F., Eyangoh S., Fontanet A. & Texier G. (2014). Spatio-temporal Patterns and Landscape-Associated Risk of Buruli Ulcer in Akonolinga, Cameroon. *PLoS Neglected Tropical Diseases*, 8, e3123.
- Leibold M.A., Holyoak M., Moquet N., Amarasekare P., Chase J.M., Hoopes M.F., Holt R.D., Shurin J.B., Law R., Tilman D., Loreau M. & Gonzalez A. (2004). The metacommunity concept: a framework for multi-scale community ecology. *Ecology Letters*, 7, 601-613.
- Loreau M. (2000). Are communities saturated? On the relationship between  $\alpha$ ,  $\beta$  and  $\gamma$  diversity. *Ecology Letters*, 3, 73-76.

- Marsollier L., Aubry J., Saint-André J.P., Robert R., Legras P., Manceau A.L., Bourdon S., Audrain C. & Carbonnelle B. (2003). Écologie et mode de transmission de *Mycobacterium ulcerans*. *Pathologie Biologie*, 51, 490-495.
- Marsollier L., Sévérin T., Aubry J., Merritt R.W., André J.-P.S., Legras P., Manceau A.-L., Chauty A., Carbonnelle B. & Cole S.T. (2004). Aquatic Snails, Passive Hosts of *Mycobacterium ulcerans*. *APPLIED AND ENVIRONMENTAL MICROBIOLOGY*, 70, 6296-6298
- Merritt R.W., Walker E.D., Small P.L.C., Wallace J.R., Johnson P.D.R., Benbow E. & Boakye D.A. (2010). Ecology and Transmission of Buruli Ulcer Disease: A Systematic Review. *PLoS Neglected Tropical Diseases*, 4, e911.
- Mihaljevic J.R. (2012). Linking metacommunity theory and symbiont evolutionary ecology. *Trends in Ecology & Evolution*, 27, 323-329.
- Moritz C., Meynard C.N., Devictor V., Guizien K., Labrune C., Guarani J.M. & Mouquet N. (2013). Disentangling the role of connectivity, environmental filtering, and spatial structure on metacommunity dynamics. *Oikos*, 122, 1401-1410.
- Morris A., Gozlan R., Marion E., Marsollier L., Andreou D., Sanhueza D., Ruffine R., Couppié P. & Guégan J.-F. (2014). First Detection of *Mycobacterium ulcerans* DNA in Environmental Samples from South America. *PLoS Neglected Tropical Diseases*, 8, e2660.
- Ostfeld R.S. (2013). A Candide response to Panglossian accusations by Randolph and Dobson: biodiversity buffers disease. *Parasitology*, 140, 1196-1198.
- Ostfeld R.S., Keesing F. & Eviner V.T. (2008). *Infectious Disease Ecology: Effects of Ecosystems on Disease and of Disease on Ecosystems*. Princeton University Press.
- Ostfeld R.S. & LoGiudice K. (2003). Community disassembly, biodiversity loss, and the erosion of an ecosystem service. *Ecology*, 84, 1421-1427.
- Portaels F., Chemlal K., Elsen P., Johnson P.D., Hayman J.A., Hibble J., Kirkwood R. & Meyers W.M. (2001). *Mycobacterium ulcerans* in wild animals. *Rev Sci Tech.*, 20, 252-264.
- R-Development-Core-Team (2011). R: A language and environment for statistical computing. In. R Foundation for Statistical Computing Vienna, Austria.
- Rafaelli D. (2004). How extinction patterns affect ecosystems. *Science*, 306, 1141-1142.
- Randolph S.E. & Dobson A.D.M. (2012). Pangloss revisited: a critique of the dilution effect and the biodiversity-buffer-disease paradigm. *Parasitology*, 139, 847-863.
- Ricklefs R.E. (1987). community diversity: relative roles of local and regional processes. *Science*, 235, 167-171.

- Roche B., Benbow M.E., Merritt R., Kimbirauskas R., McIntosh M., Small P.L.C., Williamson H. & Guegan J.-F. (2013a). Identifying the Achilles heel of multi-host pathogens: the concept of keystone 'host' species illustrated by *Mycobacterium ulcerans* transmission. *Environmental Research Letters*, 8, 045009 (7pp).
- Roche B., Rohani P., Dobson A.P. & Guegan J.-F. (2013b). The Impact of Community Organization on Vector-Borne Pathogens. *American Naturalist*, 181, 1-11.
- Shannon C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379-423.
- Stinear T.P., Seemann T., Pidot S., Frigui W., Reysset G., Garnier T., Meurice G., Simon D., Bouchier C., Ma L., Tichit M., Porter J.L., Ryan J., Johnson P.D.R., Davies J.K., Jenkin G.A., Small P.L.C., Jones L.M., Tekaia F., Laval F., Daffé M., Parkhill J. & Cole S.T. (2007). Reductive evolution and niche adaptation inferred from the genome of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer. *Genome Research*, 17, 192-200.
- Suzán G., Marcé E., Giermakowski J.T., Mills J.N., Ceballos G., Ostfeld R.S., Armién B., Pascale J.M. & Yates T.L. (2009). Experimental evidence for reduced rodent diversity causing increased hantavirus prevalence. *PLoS ONE*, 4, e5461.
- Tokeshi M. (1999). *Species coexistence. Ecological and evolutionary perspectives*. Blackwell Science Ltd, Cambridge, UK.
- Ulrich W. & Almeida-Neto M. (2012). On the meaning of nestedness: back to the basics. *Ecography*, 35, 001-007.
- Walsh D., Portaels F. & Meyers W. (2008). Buruli ulcer (*Mycobacterium ulcerans* infection). *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 102, 969-978.
- Wansbrough-Jones M. & Phillips R. (2006). Buruli ulcer: emerging from obscurity. *Lancet*, 367, 1849-1858.
- Whittaker R.H. (1972). Evolution and measurement of species diversity. *Taxon*, 21, 213-251.
- Willson S.J., Kaufman M.G., Merritt R.W., Williamson H.R., Malakauskas D.M. & Benbow M.E. (2013). Fish and amphibians as potential reservoirs of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer disease. *Infection Ecology and Epidemiology*, 3, 19946.

# Spatio-temporal Patterns and Landscape-Associated Risk of Buruli Ulcer in Akonolinga, Cameroon

Jordi Landier<sup>1,2\*</sup>, Jean Gaudart<sup>3</sup>, Kevin Carolan<sup>4</sup>, Danny Lo Seen<sup>5</sup>, Jean-François Guégan<sup>4</sup>, Sara Eyangoh<sup>6</sup>, Arnaud Fontanet<sup>1,7</sup>, Gaëtan Texier<sup>2,3</sup>

**1** Institut Pasteur, Unité de Recherche et d'expertise en Epidémiologie des Maladies Emergentes, Paris, France, **2** Service d'Epidémiologie et de Santé Publique, Centre Pasteur du Cameroun, Réseau International des Instituts Pasteur, Yaoundé, Cameroon, **3** Aix-Marseille Université, UMR912 SESSTIM (INSERM - IRD - AMU), Marseille, France, **4** UMR MIVEGEC 5290 CNRS - IRD - Université de Montpellier I - Université de Montpellier II, Montpellier, France, **5** UMR TETIS, CIRAD, Montpellier, France, **6** Service de Mycobactériologie, Centre Pasteur du Cameroun, Réseau International des Instituts Pasteur, Yaoundé, Cameroon, **7** Chaire Santé et Développement, Conservatoire National des Arts et Métiers, Paris, France



## Abstract

**Background:** Buruli ulcer (BU) is an extensively damaging skin infection caused by *Mycobacterium ulcerans*, whose transmission mode is still unknown. The focal distribution of BU and the absence of interpersonal transmission suggest a major role of environmental factors, which remain unidentified. This study provides the first description of the spatio-temporal variations of BU in an endemic African region, in Akonolinga, Cameroon. We quantify landscape-associated risk of BU, and reveal local patterns of endemicity.

**Methodology/Principal Findings:** From January 2002 to May 2012, 787 new BU cases were recorded in 154 villages of the district of Akonolinga. Incidence per village ranged from 0 (n = 59 villages) to 10.4 cases/1000 person.years (py); median incidence was 0.4 cases/1,000py. Villages neighbouring the Nyong River flood plain near Akonolinga town were identified as the highest risk zone using the SPODT algorithm. We found a decreasing risk with increasing distance to the Nyong and identified 4 time phases with changes in spatial distribution. We classified the villages into 8 groups according to landscape characteristics using principal component analysis and hierarchical clustering. We estimated the incidence ratio (IR) associated with each landscape using a generalised linear model. BU risk was highest in landscapes with abundant wetlands, especially cultivated ones (IR = 15.7, 95% confidence interval [95%CI] = 15.7[4.2–59.2]), and lowest in reference landscape where primary and secondary forest cover was abundant. In intermediate-risk landscapes, risk decreased with agriculture pressure (from IR[95%CI] = 7.9[2.2–28.8] to 2.0[0.6–6.6]). We identified landscapes where endemicity was stable and landscapes where incidence increased with time.

**Conclusion/Significance:** Our study on the largest series of BU cases recorded in a single endemic region illustrates the local evolution of BU and identifies the Nyong River as the major driver of BU incidence. Local differences along the river are explained by wetland abundance and human modification of the environment.

**Citation:** Landier J, Gaudart J, Carolan K, Lo Seen D, Guégan J-F, et al. (2014) Spatio-temporal Patterns and Landscape-Associated Risk of Buruli Ulcer in Akonolinga, Cameroon. *PLoS Negl Trop Dis* 8(9): e3123. doi:10.1371/journal.pntd.0003123

**Editor:** Mathieu Picardeau, Institut Pasteur, France

**Received:** March 26, 2014; **Accepted:** July 17, 2014; **Published:** September 4, 2014

**Copyright:** © 2014 Landier et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability:** The authors confirm that, for approved reasons, some access restrictions apply to the data underlying the findings. Data are available on demand at the National Reference Laboratory for BU in Cameroon at the Centre Pasteur du Cameroun. These data belong to the Cameroonian ministry of Health of which the Centre Pasteur du Cameroun belongs. The point of contact is Ms Valérie Flore Donfack Donkeng at LNR PNLUB, Centre Pasteur du Cameroun (+237 22 23 10 15).

**Funding:** This research was supported by a grant from the French National Research Agency ANR 11 CEPL 007 04 EXTRA-MU (<http://www.agence-nationale-recherche.fr/>). Additional funding was obtained from the Fondation Raoul Follereau (<http://www.raoul-follereau.org/>). JL is supported by fellowships from the French Ministry of Higher Education and Research and from Fondation pour la Recherche Médicale (FDT 20130928241). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* Email: jordi.landier@gmail.com

## Introduction

Buruli ulcer (BU) disease is an extensively damaging skin infection caused by *Mycobacterium ulcerans* (MU), a pathogen distantly related to *Mycobacterium tuberculosis* and *M. leprae* [1,2]. BU presents as a necrotising infection of the skin, causing debility and crippling deformity if left untreated. Initially described in Uganda and Australia [3], BU has been reported in 33 countries and mainly prevalent in tropical regions. In 2011, 4,009 cases were

reported to the World Health Organization (WHO) by 14 countries [4]. The majority of BU cases (96%) originated from countries around the gulf of Guinea and Cameroon reported 256 cases.

Means of preventing the infection are still lacking as the mode of transmission of MU to humans remains unknown [3,5]. It is unclear where the microbe resides in the environment: genomics data suggest recent adaptation to a new environmental niche and specialisation to a given host [6,7] while environmental detection



## Author Summary

Buruli ulcer (BU) remains a mysterious disease without efficient prevention since the mode of transmission of its agent, *Mycobacterium ulcerans*, is still unknown. The disease is highly localised within countries and even at the village scale within endemic regions, but environmental factors explaining this focal distribution have not been described yet. In this article, we rely on a large series of cases originating from Akonolinga region, Centre region, Cameroon, and recorded at the BU treatment center of the hospital of Akonolinga. The series of 787 patients over 10 years allows us to describe the distribution of BU incidence in the region and its changes over time and space. We identify the Nyong River as a major risk factor, and identify environmental factors along the river that further increase the risk of BU, such as the high proportion of swamps, the degradation of forests and cultivation of lands by human populations. These results will help to locate where the transmission is most likely to happen, and provide useful elements for targeting case search, prevention actions and future research on *M. ulcerans* transmission.

studies across a wide diversity of samples (insect and vertebrate fauna, water filtrates, vegetal debris) seem to indicate that the microbe could occupy a wide diversity of environments [5,8–10]. How the microbe reaches human skin is also highly debated: a role of insects as vectors has been hypothesised [9,11], but remains controversial [5].

In this context of unknown transmission mechanisms and unknown environmental location of the pathogen, the study of spatial and temporal variations of BU incidence could identify where and when MU transmission events are most likely to occur and provide insights towards understanding the elusive epidemiology of BU [12,13]. BU incidence has been described as highly focal in countries such as Uganda, Benin, Côte d'Ivoire and Ghana, where endemic regions are usually well defined. In Cameroon, three foci are identified, the Nyong valley in Ayos and Akonolinga (Centre region), Bankim (Adamaoua region) and Mbongue (South-West region), but sporadic cases originate from various places across the country, suggesting that the description of BU endemic regions is incomplete [14]. Descriptive maps of BU incidence or prevalence were established in Uganda as early as the 1970s [15,16]. Since that period, several maps of prevalence or incidence rate have been established in endemic regions from Cameroon [17–19], Democratic Republic of Congo [20], Benin [21], Côte d'Ivoire [22,23] and Ghana [24,25]. These maps showed that the distribution of BU is highly focal at country scale, and also within endemic regions [16,17,19].

Several studies have shed light on spatial patterns of BU distribution and environments associated to BU at the national scale [21,23,26] or at the regional scale [27,28]. The environmental factors associated to BU prevalence or incidence were: low elevation in Benin [21,27] and South Australia [28], high percentage of forest cover and low percentage of urban cover in Côte d'Ivoire [23], Benin [21] and South Australia [28]. In Benin, the standard deviation of wetness index, an indicator of areas with contrasted topographic features, was associated with a higher risk of BU [21]. In Côte d'Ivoire, irrigated rice producing areas and the proximity of remnant rainforest patches were associated with a higher risk of BU [23]. These analyses contributed to identifying the characteristics of the regions at risk within countries, and a model from Benin could even be used to predict where these regions would be in neighbouring Ghana [26]. However, they

provided little insight on the local determinants of BU prevalence within endemic regions, where endemic and non-endemic villages can be very close. The spatial resolution was probably too low in these studies to distinguish local variations in prevalence or to provide sufficient contrasts in the descriptions of the environment [26,28].

In this article, we present the first analysis of BU incidence patterns at the village level in an African endemic region, the Akonolinga health district, Centre region, Cameroon. This analysis is based on one of the largest series of cases available to date and on a fine scale characterisation of the environment. The objectives of this study were: 1) to describe the local spatial patterns and spatio-temporal variations of BU incidence; 2) to characterise and quantify the environmental factors associated with high BU incidence in Akonolinga district.

## Materials and Methods

### Setting

This study was performed in Akonolinga health district, located 100 km east from Yaoundé in the Centre Province of Cameroon. The predominant environment is tropical forest and the district is crossed by the Nyong River which flows from east to west. BU was first described in this area in 1977 [29]. BU prevalence in the district was described in 2001 [17] and 2007 [18]. Risk factors for BU have also been investigated in a case control study in 2006 and several individual risk factors related to the environment were identified, such as having activities in the Nyong River, or having forest or a cacao plantation close to the habitation [30]. The present study was conducted based on BU cases recorded for Disease Surveillance activities from January 2002 (start of the treatment intervention) to May 2012.

### BU case data

This study relied on the analysis of the registry [29] of BU patients included in the BU management intervention at Akonolinga District Hospital, for which the Centre Pasteur du Cameroun (CPC) performed biological confirmation of MU infection diagnosis as the National Reference Laboratory. All new patients treated for BU in Akonolinga after clinical diagnosis, and with a documented place of residence in Akonolinga district, were included in the analysis. Patients without a documented village of residence or with an unidentified village of residence were excluded. These data were collected routinely at Akonolinga District Hospital and at Centre Pasteur du Cameroun as part of the BU Disease Surveillance system of the National Control Program.

A clinical case was defined as a patient with a clinical diagnosis of BU, made at the Akonolinga District Hospital, by trained specialized health practitioners in charge of the BU treatment. A confirmed case was defined as a clinical case with a positive result for at least one of the two biological confirmation methods, microscopy [31] or PCR [32] which are performed routinely by the CPC as the National Reference Laboratory according to WHO recommendation [31]. Laboratory confirmation could not always be obtained, however clinical diagnosis was shown to be very reliable in endemic regions [33].

### Ethics statement

This study used anonymised case data, aggregated by village and by month, which were collected by the Service de Mycobactériologie of the Centre Pasteur du Cameroun as part of the surveillance activity of the National Reference Laboratory for BU in Cameroon, within the National BU Control Program.

In this study, no intervention was performed (either diagnostic or therapeutic) and we only relied on a retrospective collection of anonymous cases authorized by the Cameroonian Ministry of Health.

### Population data

Population data for Akonolinga health district villages were obtained from the national population census bureau (BUCREP). These population data were a 2010 projection based on the detailed results of the 2005 census. Population settlements in the Centre region are typically hamlets relatively close to each other which form a village (“chefferie”) under the administrative authority of a traditional chief [34,35]. Since cases were reported at the village level, we aggregated the population of hamlets at village level.

For the towns of Akonolinga and Endom, the urban neighbourhoods were aggregated.

### Administrative data

In accordance with Cameroonian laws (decree 77/245), a village was defined as the collection of all hamlets under the jurisdiction of the same traditional chief and was represented on the maps as the surface encompassing all hamlets. Hamlets had been either geolocated using a GPS during previous fieldwork [30,36] or identified on a 1/2,00 000 scale map (Institut National de Cartographie, Yaoundé, sheets of Yaoundé, Nanga Eboko and Akonolinga).

Topographical and environmental data were extracted using a circular 5 km-radius buffer around the village centroid. This value of 5 km was chosen based on a socio-anthropologic evaluation done in the region (described in [37]) and it approximated the distance that could easily be walked by inhabitants for their daily activities: fishing, farming, going to school. Furthermore, each hamlet was located within the 5-km-radius buffer of its village.

### Topographical data

Topography has been shown to be a major driver in most studies [21,23,27]. A digital elevation model was used to obtain elevation data (Shuttle Radar Topography Mission, available from the U.S. Geological Survey). A map of 90 m-topographic wetness index (TWI), an indicator of zones where water tends to accumulate due to abundant runoff from the surrounding area and a low slope, was obtained from Africa Soil Information Service (<http://www.africasoils.net>). TWI was categorised around the value 18, since following fieldwork,  $TWI > 18$  corresponded best to the bottoms of valleys which were most likely to represent wetlands. The percentage of each buffer area within this class was used as an indicator of the abundance of wetlands. Data on the distribution of rivers and roads were obtained from IFORA project and Institut National de Cartographie du Cameroun.

### Environmental data

First, we used aggregated measures to quantify vegetation cover in each buffer. We used a vegetation index calculated from remote sensing multispectral data measured by the MODIS satellite. The Enhanced Vegetation Index (EVI) is available from U.S. Geological Survey as a monthly image with 1 km<sup>2</sup> resolution averaging measurements performed with a 16-day period (30-Day L3 Global 1 km product - MOD13A3). EVI was used as a measurement of overall forest cover: it is directly related to photosynthetic activity and biomass and was developed specifically for high biomass areas such as tropical forests [38]. Using images from December 2001 to December 2011, we calculated the mean

EVI during the dry season for each village buffer to approximate vegetation cover. During the dry season, contrast was expected to be maximal between herbaceous or cultivated zones, where annual plants require rain for their growth, and forest where perennial vegetation relying on deep soil water, would still present a high photosynthetic activity [39]. Deforestation was one of the major human-driven changes that we expected in this area. We calculated the mean EVI over months December to February at the beginning and end of the study period (dry seasons 2001–2002 and 2011–12). The difference between the two values was included as a crude proxy for quantitative vegetation change over the study period.

Second, we characterised the environment in more detail (forest type, cultivated areas...) using two distinct Land-use/Land-cover (LULC) datasets.

The first dataset was a classification constructed using two Landsat images from February and March 2001 which were selected for low cloud content. Initial exploratory maps were classified using multi-spectral decision trees in the software ENVI, version 4.8 (Exelis Visual Information Solutions, Boulder, Colorado). Following ground truthing of these initial maps in November 2012, they were refined using object orientated image analysis in the software eCognition (eCognition Developer version 8.9.1, Trimble Geospatial Imaging, Munich, Germany). This resulted in regions classified as Urban, Road, Forest, Crop, Flood plain or Swamp categories.

The second dataset was a map of forest types established in 2002 and obtained from the Forest Atlas of Cameroon [40]. The study area presented 9 classes of vegetation: primary forest (dense humid evergreen or with raffia trees), secondary forest (young or adult, cultivated or not), forested wetlands/swamps, wetlands, and savannah. Secondary forest represents forest growing after being cleared (completely or partially). Two categories are distinguished according to the time elapsed since clearing. Young secondary forest corresponds to the first 5 to 20 years after clearing. It hosts mainly plant species that grow rapidly and in the light. With time, the number and variety of plant species increase, the canopy closes and the forest becomes adult secondary forest, characterised by a high biodiversity. In Akonolinga region, clearing resulted mainly from familial agriculture. The forest category indicated the intensity of human pressure on the environment.

The first dataset was used mainly for urban, agricultural, and wetland land-cover characterisation, which were combined with detailed forest data from the second dataset. New classes or new attributions were derived, such as “cultivated wetlands” corresponding to areas listed as cultivated in one dataset and swamp or swamp forest in the other.

### Statistical analyses

All analyses were performed using R software version 3.0.2 (R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria), including packages DCluster, SPODT, FactoMineR, bcp; and the software ArcGIS version 10.0 (ESRI Inc. Redlands, CA), including the extension Spatial analyst. Graphics were drawn using the ggplot2 R-package and maps were drawn using ArcGIS.

**Incidence rate calculation and mapping.** In order to analyse the distribution of cases in Akonolinga district, a map of the mean monthly incidence rate of BU per village was drawn for each phase and for the cumulative series. Mean incidence rate was expressed in cases per 1,000 person.years (cases/1,000py) and allowed comparisons between villages with different population sizes and different exposure times. To be represented on the maps, incidence rate were discretised using rounded values of the classes

obtained by the Jenks method in ArcGIS, which enabled maximization of contrasts.

**Analysis of spatial clustering.** To address the question of whether cases occurred at random in the district or according to a given spatial pattern, we performed several statistical analyses. First, a general statistic of global aggregation, Moran's Index (I), was calculated to assess spatial autocorrelation [41]. Statistical significance was calculated using bootstrap methods. Then, we evaluated the relative risk of BU over the Akonolinga region using the Spatial Oblique Decision Tree algorithm (R package SPODT). This method was used to identify homogeneous risk areas on the time-aggregated series and to quantify the risk associated with each zone. This method is adapted from classification and regression tree techniques and uses straight lines to split the study area in groups of villages as homogenous as possible regarding incidence rate [42]. It identifies clusters without any shape assumption, and is less biased by edge effects. It also provides risk estimates in all areas. Statistical significance was calculated using Monte-Carlo inference. We estimated the relative risk for each zone delimited by SPODT, by calculating an odds-ratio and its 95% significance interval.

**Spatio-temporal analysis.** Based on the spatial analysis, we analysed the incidence distribution over time and space using a "heat-map", displaying mean monthly incidence for each quarter and for each village after ordering them according to their distance to the Nyong River. We identified several "phases" in the time-series defined as periods of time presenting heterogeneous spatial patterns of incidence. These phases were confirmed using Bayesian change-point detection methods [43] to have a high probability of representing a change in the time series. Maps of incidence were drawn.

**Classification of villages into landscape groups.** We performed a principal component analysis (PCA) on the environmental data extracted for each village on a 5 km-radius buffers (see Supplementary Figure S2 legend in Text S1 for details). This step allowed grouping variables from the different categories, removing colinearity and selecting the most relevant variables for describing the environment in the Akonolinga region. Homogeneous groups of villages with similar landscape environments were built by classifying the villages according to the PCA results using agglomerative hierarchical clustering with a Euclidean distance metric (unsupervised classification).

**Estimation of landscape-associated risk.** In order to estimate the risk associated to each landscape, a generalised linear model (GLM) was built. A binomial negative regression model was preferred, since it was more adapted to this series with count numbers, as in previous BU studies [21]. Categorical variables were included in the model: landscape profile as a single variable with one modality per group, and distance to the Nyong River in 4 categories,  $\leq 5$  km,  $>5-10$  km,  $>10-20$  km,  $>20$  km, according to the activity range of populations. The model allowed estimation of an incidence ratio (IR) associated to each class. Interaction between landscape profile and distance to the Nyong River was investigated by splitting the landscape classes in groups of distance when all the villages of one landscape were not included in the same distance class. When it was found significant for one landscape, two subsets landscape were created for the final analysis, distance to Nyong  $\leq 10$  km and  $>10$  km based on the distance where Nyong influence was significant.

Univariate and multivariate models were assessed for parsimony using Akaike information criterion (AIC). Fitting was assessed by the percentage of deviance explained.

To assess model performance at representing the spatial variations of BU incidence we mapped the model residuals and

explored their distribution using Moran's I statistic. We expected that no autocorrelation would remain if the model accurately captured the spatial pattern of incidence resulting from the different landscapes.

**Temporal evolution of BU incidence in each landscape.** We drew the cumulative incidence graphs over time for each group of villages from the same landscape to examine the local temporal variations of BU. The cumulative incidence over time was fitted with a linear model. When a linear fit was appropriate and indicated constant incidence, average incidence in the landscape was calculated for the period 2002–2012. Exponential fit was also tested by fitting a curvilinear model to the logarithm of cumulative incidence.

## Results

### BU case data

From January 2002 to May 2012, 915 patients originating from Akonolinga health district were diagnosed and treated free of charge at the Akonolinga district hospital by the BU management intervention. Out of these patients, 853 were new cases and among those, 787 cases had a documented place of residence in one of the 154 villages of Akonolinga district included in the analysis. The total population of the study villages was 60,188 inhabitants and the study area had a surface of 3,685 km<sup>2</sup>. The north part of the district, 16 villages totalling 10 cases and 2,750 inhabitants, was excluded because the area was only documented in the forest LULC database.

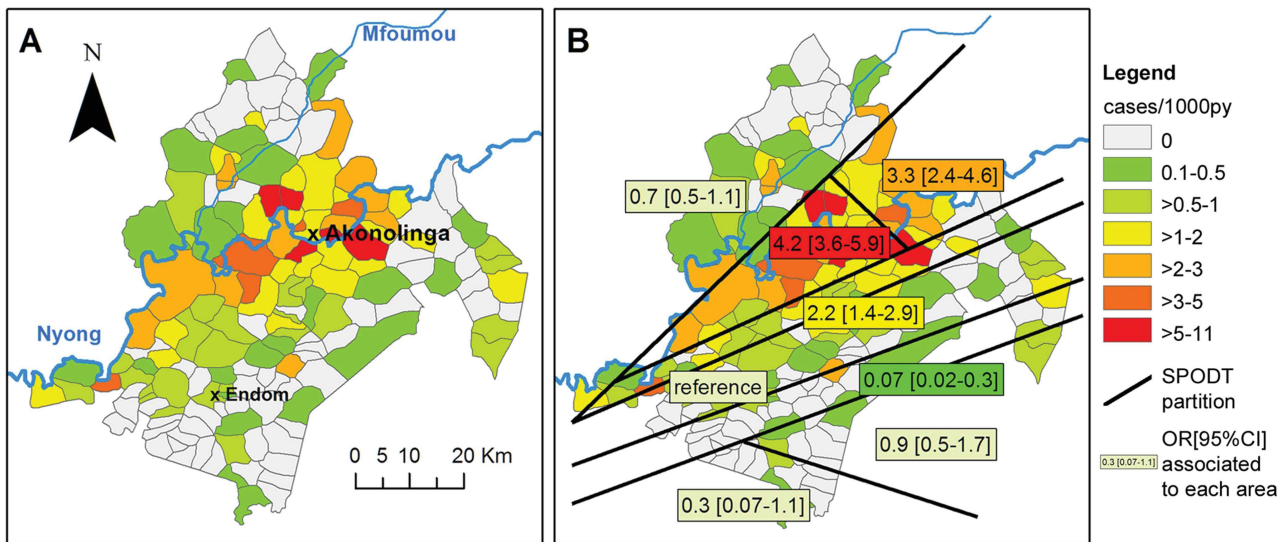
Among the 787 clinical BU cases in the database, 513 (65%) had received a laboratory-confirmed MU infection diagnosis (396 by PCR and/or microscopy, 117 by microscopy only). All 787 clinical BU cases were included in the analysis.

### BU incidence rate and spatial distribution in Akonolinga district

Global BU incidence rate in the study area was 1.25 cases/1,000py over a time period of 10 years and 5 months. Incidence per village ranged from 0, in 59 villages, to 10.4 cases/1,000py, and median incidence was 0.4 cases/1,000py (Interquartile Range = [0–1.1]). A map of cumulative incidence rate over the time-aggregated series is presented in Figure 1A. Cumulative incidence appeared to be highest in villages close to the Nyong River, east of Akonolinga town. There was a significant global aggregation of cases (Moran's  $I = 0.349$ ,  $p < 10^{-6}$ ). The SPODT algorithm identified that the highest risk zone was centered on the Nyong River upstream of Akonolinga (Figure 1B). A decreasing risk gradient with increasing distance to the Nyong was identified, and the highest risk zone had 67 times higher risk of BU than the lowest risk zone.

### Spatio-temporal variations of BU incidence in Akonolinga district, 2002–2012

The role of the Nyong River as a high risk area, and the decreasing risk gradient away from the river, led us to investigate the temporal variations of BU incidence per village according to their distance to the Nyong (Figure S1 in Supplementary Text S1). By change-point analysis process, we identified four phases corresponding to changes in the disease spatial distribution (Figure 2). In the first phase, corresponding to year 2002, the debuting BU treatment program only recruited cases from Akonolinga town and the neighbouring villages (Phase 1, Figure 2A). In the following phases, recruitment was on the entire district. From being centred on Akonolinga town from 2002 to 2006 (Phase 2, Figure 2B), the high incidence area appeared to



**Figure 1. Identification of the Nyong as a major risk factor for BU incidence in Akonolinga 2002–2012 (spatial analysis on time-aggregated incidence rate of BU in Akonolinga).** A: Incidence rate per village (cases/1,000py). B: Decreasing risk of BU with increasing distance to the Nyong River. Homogenous risk areas of Akonolinga district were identified using the SPODT algorithm. Associated odds-ratio and 95% CI are provided.

doi:10.1371/journal.pntd.0003123.g001

move, first eastward upstream the Nyong in the area of Abem (Phase 3, Figure 2C), then downstream along the Nyong, on the southern part of the river and on the Mfoumou, a tributary of the Nyong (Phase 4, Figure 2D).

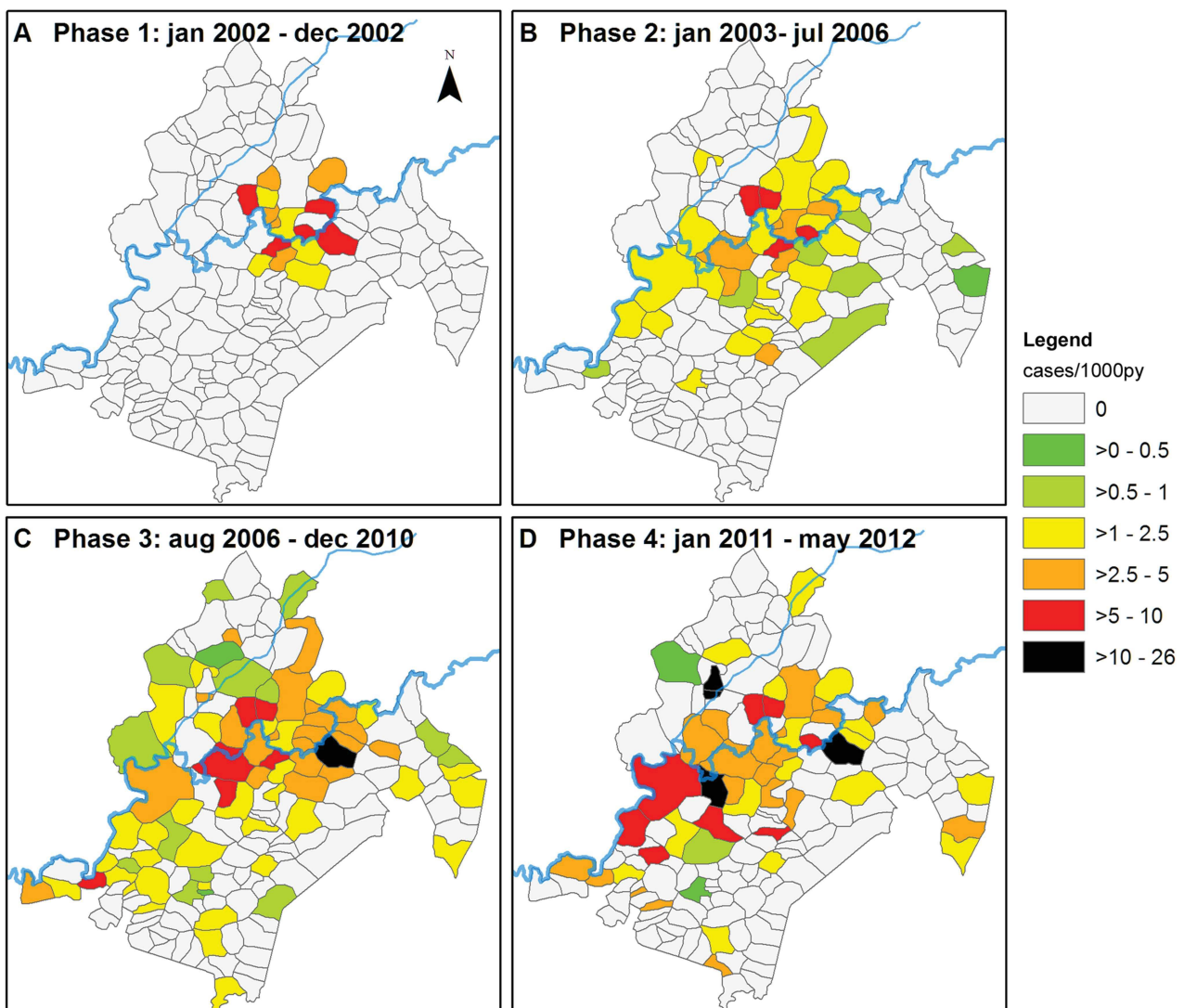
#### Analysis of local environmental risk-factors for BU in Akonolinga district

**Environment classification in landscape groups.** The unsupervised classification of villages distinguished 7 landscape groups, organised on two main gradients (Table 1 and details in Supplementary Text S1, Figure S2). First, a clear separation was observed between villages with abundant forest cover compared to those where it was greatly reduced, as indicated by EVI values. This separation allowed the definition of a first gradient of increasing human alteration of landscape, based on abundant urban and agricultural land-use, and low proportion of forest cover. Landscape “Urban Nyong” and landscape “Rural Nyong” were characterised by a low forest cover and abundant areas dedicated to agriculture, as well as a high proportion of wetlands. A second gradient separated the villages according to the forest maturity (primary, secondary adult, secondary young) and the proportion of which was mosaicked with cultures. Landscapes “Forest 1” and “Forest 2” had the most abundant forest cover and were generally at a higher altitude with lower proportion of wetlands. “Forest 1” included remnants of dense humid evergreen primary forest, which marks the persistence of undisturbed ecosystems. “Forest 2” included a high proportion of secondary adult forest cover, a fraction of which was cultivated. Landscapes “Cultivated forest” and “Young forest” included intermediate features between these two groups, where young secondary forest, cultivated or not, dominated, indicating a more intense agricultural pressure. Both also presented abundant proportion of wetlands. Finally, landscape “Savannah” corresponded to 2 villages located in a specific area of savannah within the forest. Changes in forest cover, approximated by EVI difference between dry season 2001–02 and 2011–2012 were highest in landscapes “Forest 2” and “Young forest”.

**Estimation of landscape-associated risk.** The use of a generalised linear model allowed estimation of BU incidence ratio (IR) from January 2002 to May 2012. Univariate analysis is presented in Table 2. In the landscape model, the highest risk zones corresponded to landscapes “Urban Nyong” and “Rural Nyong” compared to landscape “Forest 1”. “Young forest” presented an intermediate risk and all other landscapes did not significantly differ from “Forest 1”. In the Nyong River distance model, risk decreased with increasing distance to the river with a dose-response relationship.

For multivariate analysis, we combined Nyong distance and landscape and split landscape “Cultivated forest” according to the location of villages within or beyond the influence range of the Nyong, i.e. “Cultivated forest,  $\leq 10$  km from Nyong” and “Cultivated forest,  $> 10$  km from Nyong”. All other landscapes were located within a single Nyong distance class or did not present significant differences in IR between the two distance classes. The resulting model (Table 3 and figure 3A) indicated that “Urban Nyong” and “Rural Nyong” had the highest risk, respectively IR = 15.7 (95%CI = [4.2–59.2]) and IR = 12.5 (95%CI = [3.7–42.8]) compared to landscape “Forest 1”. Landscapes “Young forest” and “Cultivated forest  $\leq 10$  km from Nyong” had intermediate risk, respectively IR = 7.9 (95%CI = [2.2–28.8]) and IR = 4.9 (95%CI = [1.4–17.4]). Finally, risk for landscapes “Forest 2”, “Cultivated forest  $> 10$  km from Nyong” and “Savannah” did not significantly differ from the “Forest 1” landscape. This model explained 41% of the variance between the villages. We performed a further analysis on the model residuals, and found that their spatial distribution presented no remaining spatial autocorrelation (Moran’s I = 0.021,  $p = 0.65$ ). This indicated that our model was able to capture most of the spatial pattern between the villages. Predicted incidence rate and actual cumulative incidence rate maps are presented in Figure 3B and 3C.

**Temporal variations of BU incidence in each landscape.** We studied the series of monthly incident cases for each landscape in order to characterise the temporal variations of



**Figure 2. Maps of spatio-temporal variations of BU incidence in Akonolinga district.** A–D: Incidence rate maps for the periods, phases 1 to 4, identified in the time-series (cases/1,000py). doi:10.1371/journal.pntd.0003123.g002

BU incidence within the different landscapes (Supplementary Text S1, Figure S3). Landscape “Forest 1” presented only 4 cases during the study period (incidence of 0.2 cases/1,000py) and “Savannah” only 3 cases (0.5 cases/1,000py). Landscapes “Urban Nyong”, “Rural Nyong”, and “Forest 2” presented stable incidence rates over the study period, averaging respectively 2.1, 2.4 and 0.4 cases/1,000py. Finally, incidence was increasing in landscapes “Cultivated forest” and “Young forest”. “Cultivated forest  $\leq 10$  km from Nyong” even presented an exponentially increasing incidence rate ( $R^2 = 0.97$  for exponential fit compared to  $R^2 = 0.87$  for linear fit).

## Discussion

Our study relied on the analysis of 787 BU cases over 125 months of follow-up, which to our knowledge is amongst the highest reported incidences in an endemic region with more than 10 years of continuous follow-up [28,44]. We analysed BU spatio-temporal patterns and were able to reveal local-scale

environmental determinants of BU incidence. We demonstrated that the Nyong River represented a major risk factor for BU, in conformity with previous studies of individual risk factors [30] and environmental MU detection [36,45]. We also identified different levels of risk along the river, which were associated to different environment profiles. We suggest that BU risk further increases with abundance of wetlands and with human modifications of landscape, such as cultivation and forest clearing. We also identified stable endemic areas and zones where incidence appears to be rising.

This work benefited from several methodological improvements compared to previous studies. By using the SPODT algorithm for identification of risk zones, we obtained a more accurate description than other studies [21], showing a decreasing risk gradient away from the Nyong River. By considering different categories of forest cover and management, cultivated and uncultivated wetlands, we accounted for local heterogeneities which would have been missed in broader analyses. Contrary to previous studies, which considered forest as a homogeneous

**Table 1.** Selected environment characteristics of landscape groups defined in Akonolinga district.

Landscape	Major features*	Mean EVI in December*	EVI decrease 2001–12*	Area with WI>18*	Forested Wetland*	Cultivated Wetland*	N villages	Total Population	Main watershed
<b>Savannah</b>	Savannah (81%)	0,367	0,049	6%	0	0	2	538	Nyong
<b>Urban Nyong</b>	Young cultivated secondary forest (29%), Urban (5%), and cultivated land (9%)	0,381	0,047	37%	7%	8%	7	17813	Nyong
<b>Rural Nyong</b>	Young cultivated secondary forest (25%)	0,392	0,062	20%	5%	3%	20	6656	Nyong
<b>Cultivated Forest &lt;10 km Nyong</b>	Young cultivated secondary forest (37%)	0,417	0,039	6%	10%	3%	16	5015	Nyong
<b>Cultivated Forest &gt;10 km Nyong</b>	Young cultivated secondary forest (39%)	0,416	0,026	7%	10%	2%	31	8529	Dja
<b>Young Forest</b>	Young secondary forest (38%)	0,417	0,074	5%	9%	+	12	3069	Nyong
<b>Forest 2</b>	Adult secondary forest (50%)	0,415	0,076	5%	0	+	58	16729	Nyong
<b>Forest 1</b>	Adult secondary forest (48%) and Primary forest (18%)	0,420	0,068	4%	+	+	8	1839	Dja

\*median value for each landscape group; + present, <1%.  
doi:10.1371/journal.pntd.0003123.t001

**Table 2.** Univariate analysis.

		IRR <sup>1</sup>	[95%CI] <sup>2</sup>	p-value
<b>Landscape group</b>	Urban Nyong	15.8	4.05–61.32	<0.001
	Rural Nyong	12.6	3.6–44.0	<0.001
	Savannah	5.4	0.8–36.4	0.081
	Cultivated Forest	2.8	0.8–9.5	0.106
	Young forest	7.9	2.1–29.6	0.002
	Forest 2	2.0	0.6–6.7	0.287
	Forest 1	1	reference	
<b>Distance to the Nyong</b>	≤5 km	7.5	4.3–13.5	<0.001
	>5 and ≤10 km	2.3	1.2–4.3	0.01
	>10 and ≤20 km	1.2	0.7–2.2	0.528
	>20 km	1	reference	

Incidence rate ratios estimated for the landscape groups and the distance to the Nyong River in 154 villages of Akonolinga district, Cameroon, 2002–2012.

<sup>1</sup>IRR: Incidence Rate ratio.

<sup>2</sup>[95%CI]: 95% confidence interval.

doi:10.1371/journal.pntd.0003123.t002

category [21,23,28] and found it a risk factor, we used a detailed LULC classification, ground-truthed, and a small buffer radius (5 km instead of 20 km). We demonstrated that the different forest categories presented different risk levels according to their status regarding human activities, and that BU risk followed a dose-response relationship according to forest degradation [5,46].

Our study showed that BU incidence spatio-temporal patterns are complex, but might be explained for a large part by landscape characteristics and heterogeneities. We identified the Nyong River as a major driver of BU incidence in the Akonolinga region, and local scale environmental variations in the landscapes along the river were associated to significantly different risk levels.

These variations, distinguishing between landscapes at high and intermediate BU-risk were principally the proportion of wetlands, and the type and extent of forest cover. The proportion of wetlands was evaluated topographically (% surface with TWI>18) or in LULC descriptions, where cultivated wetlands occupied a larger surface in high-risk landscapes (“Rural Nyong” and “Urban Nyong”) and forested wetlands in intermediate-risk landscapes (“Young forest” and “Cultivated forest”). The type and extent of forest cover reflected the level of human modifications. In the

highest-risk landscapes, forest cover was reduced and corresponded mainly to cultivated young secondary forest. These landscapes, located in the densely populated part of the district, are shaped by intense agricultural pressure, as indicated also by the proportion of cultivated lands, including wetlands. The intermediate-risk landscapes near the Nyong River, “Cultivated forest <10 km from the Nyong River” and “Young forest”, were less modified by human activities and retained important forest covers. The observed increase in incidence during the study period could result from recent environmental modifications: using only a crude measurement, we showed that “Young forest” is one of the landscapes with the largest decrease in EVI, indicating a decrease in forest cover. These areas of increasing incidence are located downstream from the floodplain of the town of Akonolinga. Speculatively, MU could have spread along the Nyong colonising new environments.

The landscapes at lowest risk, “Forest 1”, “Forest 2” and “Cultivated forest >10 km from Nyong River”, were mainly composed of villages located far from the Nyong River and corresponding to the most preserved environments. In “Forest 1” landscape, BU incidence was about 100 times lower than in highest risk areas, while it was only about 50 times lower in “Forest

**Table 3.** Incidence rate ratios estimated for the landscape groups combined with Nyong River distance in 154 villages of Akonolinga health district, Cameroon, 2002–2012.

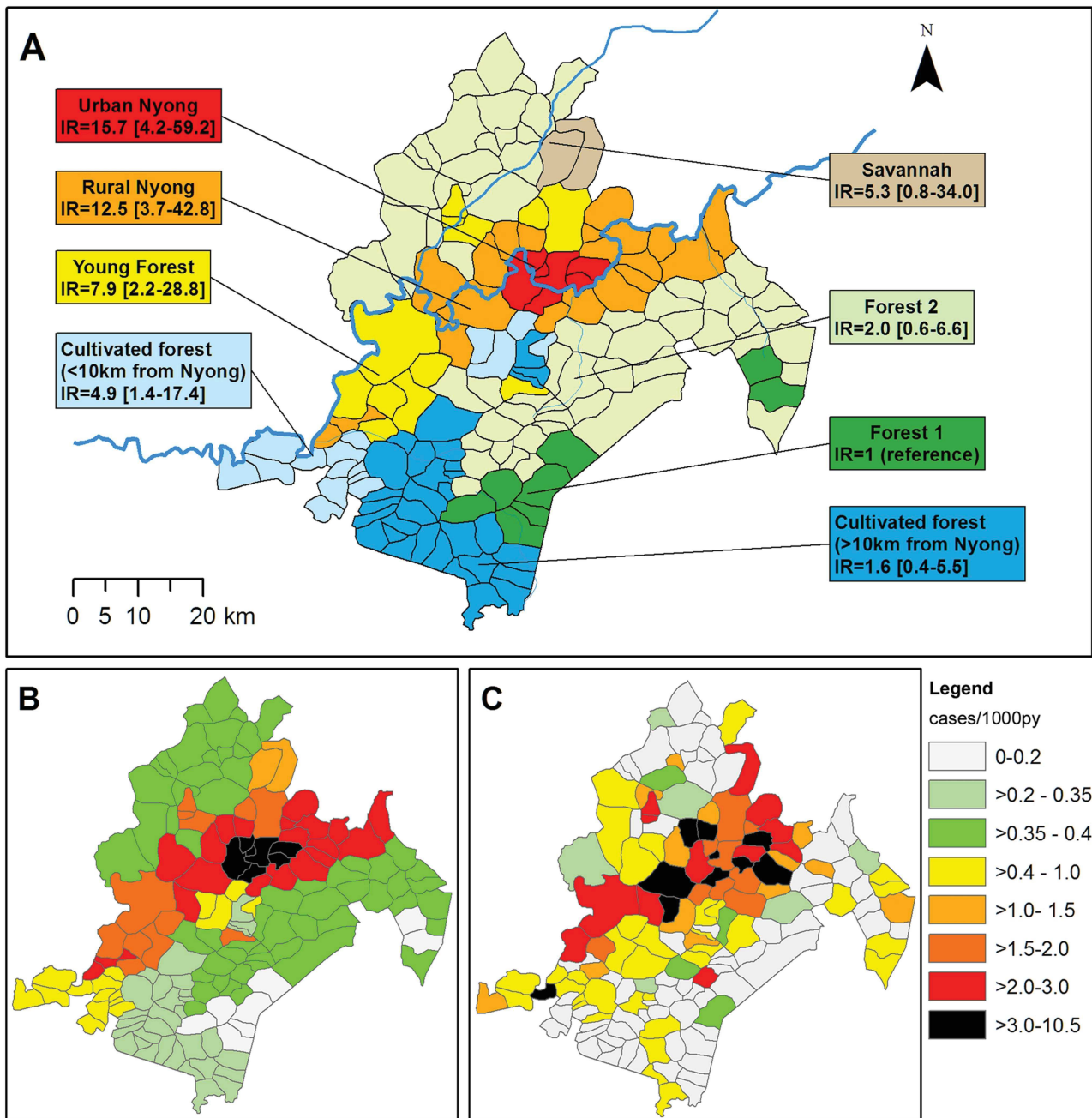
Landscape group	IRR <sup>1</sup>	[95%CI] <sup>2</sup>	p-value
Urban Nyong	15.7	4.2–59.4	<0.001
Rural Nyong	12.5	3.7–42.9	<0.001
Savannah	5.4	0.8–34.2	0.077
Cultivated forest; ≤10 km to Nyong	4.9	1.4–17.4	0.014
Cultivated forest; >10 km to Nyong	1.6	0.4–5.5	0.499
Young forest	7.9	2.2–28.9	0.002
Forest 2	2.0	0.6–6.6	0.277
Forest 1	1	reference	

Deviance explained: 41%; Akaike Information Criterion: 578.4.

<sup>1</sup>IRR: Incidence Rate ratio.

<sup>2</sup>[95%CI]: 95% confidence interval.

doi:10.1371/journal.pntd.0003123.t003



**Figure 3. Landscape-associated risk of BU in Akonolinga district, 2002-2012.** A: Classification of Akonolinga area villages according to landscape group and associated BU incidence ratio with 95% confidence interval. B: Predicted cumulative incidence for each village of the district according to the landscape model (cases/1,000py). C: Observed cumulative incidence rate for each village of the district (cases/1,000py). doi:10.1371/journal.pntd.0003123.g003

2” landscape. Even if not significant, we observed the same trend of BU risk increase with increasing forest degradation level.

We can propose that spatial variations of BU incidence in Akonolinga Health District resulted from the superimposition of two main factors: a high or low baseline risk related to the Nyong River proximity, and additional risks related to wetland abundance and environmental modifications by human activities. The role of the wetlands was supported by analyses of MU presence in

Akonolinga water bodies, which showed that wetlands acted as a permanent reservoir of MU over the year, while other water bodies presented season-specific peaks of MU colonisation [45]. The increase in BU risk associated with human modifications of the environment could result from contact with newly accessible but pre-existing high-risk environments, from an increase in the number of contacts with risk sources due to populations increasing their activity range, or from the transformation of natural



environments into high-risk sources by human activities (such as clearing wetlands for cultivation) [47]. The contribution of each phenomenon could be evaluated using chronological descriptions of the environment evolution, as well as of human practices.

The main limitation of this work was that it relied on “semi-active” case detection. We analysed data from cases which were diagnosed and treated at Akonolinga district hospital. They may not represent all the cases that occurred in the Akonolinga health district over the study period, since patients tend to seek traditional treatments as a first option [48]. However, since BU is a slowly progressing disease and difficult to cure even in hospitals, cases are likely to seek medical care at some point, eventually after failure of traditional treatment [48]. We analysed the spatial incidence trends over large periods of time, which probably allowed us to capture a large proportion of the incident cases over the study period, even patients with long delay to diagnosis. In addition, we cannot be sure that cases were infected in a given location, but given the activity patterns and our study scale, we can be confident that our work focuses on the main environments frequented by the populations.

This series of cases originated from a single treatment centre with a defined population-catchment area. The BU program in Akonolinga district has established a dynamic network of community correspondents in the villages, who contribute to population information and awareness, as well as case detection. Regular investigations are also performed in the area by medical staff and social workers in order to examine suspect cases and advocate for hospital treatment in a context where traditional treatment is generally the first option, despite free treatments [46]. This coverage ensured that a maximum number of incident cases were detected, diagnosed and treated, and therefore included in this analysis. Comparisons with previous data from cross-sectional surveys in Akonolinga region support our assumption that cases treated at Akonolinga hospital are representative of cases occurring in the Akonolinga district [17,18] and support our description of localised increases in incidence.

Our results are consistent with previous results regarding individual risk factors. In a case-control study from 2006, wading in the Nyong swamps and not wearing long protective clothing while farming were identified as risk factors [30]. We showed that BU is associated to agricultural areas near the Nyong and suggested that the clearing and cultivation of swamps could have contributed to risk increase at the population level. This hypothesis was supported by interviews performed during an anthropologic study in Ekougou and Abem, two villages located near the Nyong flood plain in “Rural Nyong” landscape. The informants related BU to the practice of clearing swamps for vegetable cultivation, which started about 20–25 years ago in the flood plain near Akonolinga [37]. It would be interesting to document if these practices occur in landscapes with increasing incidence, and when

they started. The informants also incriminated large bushfires in the 1980s, which deeply modified the ecosystem of the Nyong floodplain [37]. The increase in population could also explain a new need for land in more remote areas of the district, triggering deforestation and BU.

## Conclusion

The present work provides a quantitative assessment of the link between BU, slow flowing rivers, like the Nyong River, landscape features and their modifications by human activities. We clarify the role of forest, previously considered as a risk factor, by distinguishing pristine from human-perturbed ecosystems. We also underline major heterogeneities within Akonolinga endemic area, which presents stable high and low endemic zones, and zones with a rising incidence rate. Further studies regarding environment sampling for MU detection in endemic areas, or identification of risk-factors should take into account that environments at risk are defined at a very local scale. Surveillance of BU and active case search programs in endemic regions should also include the fact that BU geography can be substantially modified on a short time span, endangering new populations.

## Supporting Information

**Checklist S1 STROBE checklist.**  
(DOC)

**Text S1 Supporting figures and detailed legends.** Includes: Figure S1: Space-time analysis of BU incidence in Akonolinga. Figure S2: Principal components analysis and selected results. Figure S3: Cumulative incidence graphs in 6 Akonolinga landscapes, 2002–2012.  
(PDF)

## Acknowledgments

The authors would like to thank Tamara Giles-Vernick for sharing her understanding of population practices in Akonolinga, Juliette Paireau for numerous tips using R and ArcGIS, Benjamin Roche and Romulus Breban for useful insights on data analysis, Joachim Etouana (Institut National de Cartographie du Cameroun), Minia Valère (BUCREP) who contributed population and administrative data, Hervé Chevillotte (IRD Cameroon), for environmental data from the IFORA project (ANR-Biodiv grant IFORA), the partner teams of the EXTRA-MU project, the Mycobacteriology Service staff at Centre Pasteur du Cameroun who performed diagnosis confirmation, and Drs Gervais Folefack and Michael Tsoungui at Akonolinga district hospital and their staff.

## Author Contributions

Conceived and designed the experiments: JL JG GT KC JFG AF. Performed the experiments: JL GT KC. Analyzed the data: JL JG KC GT DLS. Contributed reagents/materials/analysis tools: SE DLS. Contributed to the writing of the manuscript: JL JG KC JFG DLS SE AF GT.

## References

- Silva MT, Portaels F, Pedrosa J (2009) Pathogenetic mechanisms of the intracellular parasite *Mycobacterium ulcerans* leading to Buruli ulcer. *Lancet Infect Dis* 9: 699–710. doi:10.1016/S1473-3099(09)70234-8.
- Wansbrough-Jones M, Phillips R (2006) Buruli ulcer: emerging from obscurity. *Lancet* 367: 1849–1858. doi:10.1016/S0140-6736(06)68807-7.
- Johnson PDR, Stinear T, Small PLC, Pluschke G, Merritt RW, et al. (2005) Buruli ulcer (*M. ulcerans* infection): new insights, new hope for disease control. *PLoS Med* 2: e108. doi:10.1371/journal.pmed.0020108.
- World Health Organisation - Global Buruli Ulcer Initiative Meeting (2012) Buruli ulcer control and research: Summary Data. Available: [http://www.who.int/buruli/Summary\\_data\\_and\\_new\\_target\\_2013.pdf](http://www.who.int/buruli/Summary_data_and_new_target_2013.pdf). Accessed 12 December 2013.
- Merritt RW, Walker ED, Small PLC, Wallace JR, Johnson PDR, et al. (2010) Ecology and transmission of Buruli ulcer disease: a systematic review. *PLoS Negl Trop Dis* 4: e911. doi:10.1371/journal.pntd.0000911.
- Stinear TP, Seemann T, Pidot S, Frigui W, Reyssat G, et al. (2007) Reductive evolution and niche adaptation inferred from the genome of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer. *Genome Res* 17: 192–200. doi:10.1101/gr.5942807.
- Doig KD, Holt KE, Fyfe JAM, Lavender CJ, Eddyani M, et al. (2012) On the origin of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer. *BMC Genomics* 13: 258. doi:10.1186/1471-2164-13-258.
- Portaels F, Elsen P, Guimaraes-Peres a, Fonteyne P a, Meyers WM (1999) Insects in the transmission of *Mycobacterium ulcerans* infection. *Lancet* 353: 986. doi:10.1016/S0140-6736(98)05177-0.
- Johnson PDR, Azuolas J, Lavender CJ, Wishart E, Stinear TP, et al. (2007) *Mycobacterium ulcerans* in mosquitoes captured during outbreak of Buruli ulcer, southeastern Australia. *Emerg Infect Dis* 13: 1653–1660. doi:10.3201/cid1311.061369.

10. Fyfe JAM, Lavender CJ, Handsyde KA, Legione AR, O'Brien CR, et al. (2010) A major role for mammals in the ecology of *Mycobacterium ulcerans*. *PLoS Negl Trop Dis* 4: e791. doi:10.1371/journal.pntd.0000791.
11. Marsollier L, Robert R, Aubry J, Andre J, Saint, Kouakou H, et al. (2002) Aquatic insects as a vector for *Mycobacterium ulcerans*. *Appl Environ Microbiol* 68: 4623–4628. doi:10.1128/AEM.68.9.4623.
12. Lambin EF, Tran A, Vanwambeke SO, Linard C, Soti V (2010) Pathogenic landscapes: interactions between land, people, disease vectors, and their animal hosts. *Int J Health Geogr* 9: 54. doi:10.1186/1476-072X-9-54.
13. Faucher B, Gaudart J, Faraut F, Pomares C, Mary C, et al. (2012) Heterogeneity of environments associated with transmission of visceral leishmaniasis in South-Eastern France and implication for control strategies. *PLoS Negl Trop Dis* 6: e1765. doi:10.1371/journal.pntd.0001765.
14. Um Boock A, Wiss K, Eyangoh S (2005) [Preliminary national survey on Buruli ulcer in Cameroon]. *Bull FALF*: 41–42.
15. Uganda Buruli Group (1971) Epidemiology of *Mycobacterium ulcerans* infection (Buruli ulcer) at Kinyara, Uganda. *Trans R Soc Trop Med Hyg* 65: 763–775.
16. Barker DJ (1972) The distribution of Buruli disease in Uganda. *Trans R Soc Trop Med Hyg* 66: 367–374.
17. Noeske J, Kuaban C, Rondini S, Sorlin P, Ciaffi L, et al. (2004) Buruli ulcer disease in Cameroon rediscovered. *Am J Trop Med Hyg* 70: 520–526.
18. Porten K, Sailor K, Comte E, Njikap A, Sobry A, et al. (2009) Prevalence of Buruli ulcer in Akonolinga health district, Cameroon: results of a cross sectional survey. *PLoS Negl Trop Dis* 3: e466. doi:10.1371/journal.pntd.0000466.
19. Marion E, Landier J, Boisier P, Marsollier L, Fontanet A, et al. (2011) Geographic expansion of Buruli ulcer disease, Cameroon. *Emerg Infect Dis* 17: 551–553. doi:10.3201/eid1703091859.
20. Phanzu DM, Suykerbuyk P, Saunderson P, Ngwala Lukanu P, Masamba Minuku J-B, et al. (2013) Burden of *Mycobacterium ulcerans* disease (Buruli ulcer) and the underreporting ratio in the territory of Ssongololo, Democratic Republic of Congo. *PLoS Negl Trop Dis* 7: e2563. doi:10.1371/journal.pntd.0002563.
21. Wagner T, Benbow ME, Brenden TO, Qi J, Johnson RC (2008) Buruli ulcer disease prevalence in Benin, West Africa: associations with land use/cover and the identification of disease clusters. *Int J Health Geogr* 7: 25. doi:10.1186/1476-072X-7-25.
22. Marston BJ, Diallo MO, Horsburgh CR, Diomande I, Saki MZ, et al. (1995) Emergence of Buruli ulcer disease in the Daloa region of Cote d'Ivoire. *Am J Trop Med Hyg* 52: 219–224.
23. Brou T, Broutin H, Elguero E, Asse H, Guegan J-F (2008) Landscape diversity related to Buruli ulcer disease in Côte d'Ivoire. *PLoS Negl Trop Dis* 2: e271. doi:10.1371/journal.pntd.0000271.
24. Amofah G, Bonsu F, Tetteh C, Okrah J, Asamoah K, et al. (2002) Buruli ulcer in Ghana: results of a national case search. *Emerg Infect Dis* 8: 167–170. doi:10.3201/eid0802.010119.
25. Williamson HR, Benbow ME, Nguyen KD, Beachboard DC, Kimbirauskas RK, et al. (2008) Distribution of *Mycobacterium ulcerans* in buruli ulcer endemic and non-endemic aquatic sites in Ghana. *PLoS Negl Trop Dis* 2: e205. doi:10.1371/journal.pntd.0000205.
26. Wagner T, Benbow ME, Burns M, Johnson RC, Merritt RW, et al. (2008) A Landscape-based model for predicting *Mycobacterium ulcerans* infection (Buruli ulcer disease) presence in Benin, West Africa. *Ecohealth* 5: 69–79. doi:10.1007/s10393-007-0148-7.
27. Sopath GE, Johnson RC, Anagonou SY, Barogui YT, Dossou AD, et al. (2011) Buruli ulcer prevalence and altitude, Benin. *Emerg Infect Dis* 17: 153–154. doi:10.3201/eid1701.100644.
28. Van Ravensway J, Benbow ME, Tsonis AA, Pierce SJ, Campbell LP, et al. (2012) Climate and landscape factors associated with Buruli ulcer incidence in Victoria, Australia. *PLoS One* 7: e51074. doi:10.1371/journal.pone.0051074.
29. Ravisse P (1977) [Skin ulcer caused by *Mycobacterium ulcerans* in Cameroon. I. Clinical, epidemiological and histological study]. *Bull Soc Pathol Exot Filiales* 70: 109–124.
30. Pouillot R, Matias G, Wondje CM, Portaels F, Valin N, et al. (2007) Risk factors for buruli ulcer: a case control study in Cameroon. *PLoS Negl Trop Dis* 1: e101. doi:10.1371/journal.pntd.0000101.
31. World Health Organization, Portaels F (2014) Laboratory diagnosis of Buruli ulcer. World Health Organization.
32. Fyfe JAM, Lavender CJ, Johnson PDR, Globan M, Sievers A, et al. (2007) Development and application of two multiplex real-time PCR assays for the detection of *Mycobacterium ulcerans* in clinical and environmental samples. *Appl Environ Microbiol* 73: 4733–4740. doi:10.1128/AEM.02971-06.
33. Mensah-Quainoo E, Yeboah-Manu D, Asebi C, Patafuor F, Ofori-Adjei D, et al. (2008) Diagnosis of *Mycobacterium ulcerans* infection (Buruli ulcer) at a treatment centre in Ghana: a retrospective analysis of laboratory results of clinically diagnosed cases. *Trop Med Int Health* 13: 191–198. doi:10.1111/j.1365-3156.2007.01990.x.
34. Laburthe-Tolra P (1981) Les seigneurs de la forêt. Les seigneurs de la forêt. Essai sur le passé historique, l'organisation sociale et les normes éthiques des anciens Beti du Cameroun. Paris: L'Harmattan.
35. Guyer J (1984) Family and farm in Southern Cameroon. Boston: Boston University African Studies Center.
36. Marion E, Eyangoh S, Yeramian E, Doannio J, Landier J, et al. (2010) Seasonal and regional dynamics of *M. ulcerans* transmission in environmental context: deciphering the role of water bugs as hosts and vectors. *PLoS Negl Trop Dis* 4: e731.
37. Giles-Vernick T, Owona-Ntsama J, Landier J, Eyangoh S (2014) The puzzle of Buruli ulcer transmission, ethno-ecological history and the end of "love" in the Akonolinga district, Cameroon. *Soc Sci Med*. doi:10.1016/j.socscimed.2014.03.008.
38. Gond V, Fayolle A, Pennec A, Cornu G, Mayaux P, et al. (2013) Vegetation structure and greenness in Central Africa from Modis multi-temporal data. *Philos Trans R Soc B* 368: 20120309.
39. Huete AR, Didan K, Shimabukuro YE, Ratana P, Saleska SR, et al. (2006) Amazon rainforests green-up with sunlight in dry season. *Geophys Res Lett* 33: L06405. doi:10.1029/2005GL025583.
40. World Resources Institute (2013) Interactive Forest Atlas of Cameroon (version 3.0).
41. Gaudart J, Giorgi R, Poudiougou B, Touré O, Ranque S, et al. (2007) [Spatial cluster detection without point source specification: the use of five methods and comparison of their results]. *Rev Epidemiol Sante Publique* 55: 297–306. doi:10.1016/j.respe.2007.04.003.
42. Gaudart J, Poudiougou B, Ranque S, Doumbo O (2005) Oblique decision trees for spatial pattern detection: optimal algorithm and application to malaria risk. *BMC Med Res Methodol* 5: 22. doi:10.1186/1471-2288-5-22.
43. Erdman C, Emerson JW (2007) bcp: An R package for performing a Bayesian analysis of Change Point problems. *J Stat Softw* 23: 1–13.
44. Morris A, Gozlan R, Marion E, Marsollier L, Andreou D, et al. (2014) First detection of *Mycobacterium ulcerans* DNA in environmental samples from South America. *PLoS Negl Trop Dis* 8: e2660. doi:10.1371/journal.pntd.0002660.
45. Garchitorena A, Roche B, Kamgang R, Ossomba J, Babonneau J, et al. (2014) *Mycobacterium ulcerans* ecological dynamics and its association with freshwater ecosystems and aquatic communities: results from a 12-month environmental survey in Cameroon. *PLoS Negl Trop Dis* 8: e2879. doi:10.1371/journal.pntd.0002879.
46. Hill AB (1965) The environment and disease: association or causation? *Proc R Soc Med* 58: 295–300.
47. Kilpatrick AM, Randolph SE (2012) Drivers, dynamics, and control of emerging vector-borne zoonotic diseases. *Lancet* 380: 1946–1955. doi:10.1016/S0140-6736(12)61151-9.
48. Peeters Grietens K, Toomer E, Um Boock A, Hausmann-Muela S, Peeters H, et al. (2012) What role do traditional beliefs play in treatment seeking and delay for Buruli ulcer disease?—insights from a mixed methods study in Cameroon. *PLoS One* 7: e36954. doi:10.1371/journal.pone.0036954.