

The Effect of Land Use and Climate on Malaria Risk in the Amazon Region

by

Denis Ribeiro do Valle

University Program in Ecology  
Duke University

Date: \_\_\_\_\_

Approved:

\_\_\_\_\_  
James S. Clark, Supervisor

\_\_\_\_\_  
Alan Gelfand

\_\_\_\_\_  
Subhrendu Pattanayak

\_\_\_\_\_  
William K. Pan

\_\_\_\_\_  
Katharina V. Koelle

Thesis submitted in partial fulfillment of  
the requirements for the degree of Doctor of Philosophy  
in the University Program in Ecology  
in the Graduate School  
of Duke University

2013

ABSTRACT

The Effect of Land Use and Climate on Malaria Risk in the Amazon Region

by

Denis Ribeiro do Valle

University Program in Ecology  
Duke University

Date: \_\_\_\_\_

Approved:

\_\_\_\_\_  
James S. Clark, Supervisor

\_\_\_\_\_  
Alan Gelfand

\_\_\_\_\_  
Subhrendu Pattanayak

\_\_\_\_\_  
William K. Pan

\_\_\_\_\_  
Katharina V. Koelle

An abstract of a thesis submitted in partial  
fulfillment of the requirements for the degree  
of Doctor of Philosophy in the  
University Program in Ecology  
in the Graduate School of  
Duke University

2013

Copyright by  
Denis Ribeiro do Valle  
2013

## Abstract

The goal of this thesis is to study the relationship between climate and land use / land cover (LULC) on malaria risk in the Amazon region. Despite the large public policy implications, current literature provides contradictory evidence regarding how LULC affects malaria risk. Furthermore, little is known regarding the public health impacts of the predicted drying of the Amazon. In this thesis, we rely on mosquito and malaria incidence/prevalence data from multiple sources, both from the Brazilian and Peruvian Amazon, and we develop novel methodology to integrate multiple datasets and infer malaria risk factors.

The first chapter describes a novel method to combine data from different *Plasmodium* detection methods (i.e., polymerase chain reaction (PCR) and microscopy) and sampling schemes (i.e., aggressive active case detection and passive case detection). Using this method and detailed data on malaria prevalence, we find that proximity to forest, as well as participation on forest activities, greatly enhance malaria risk. Furthermore, our results suggest that asymptomatic *Plasmodium* carriers are more likely to be individuals that have recently arrived in the region and that live in regions with high forest cover.

The second chapter describes large scale patterns regarding LULC, climate, and malaria incidence. To this end, we rely on a unique dataset collected by the Brazilian

government surveillance system over ~4 years, spanning ~4.5 million km<sup>2</sup> and 1,300,000 malaria cases. This analysis indicates that malaria incidence is substantially higher in areas with higher forest cover, whereas deforestation rate (the often cited culprit of malaria in the region) is not significantly associated with malaria. We also find that a drier climate may enhance malaria risk in the Brazilian Amazon region. We then employ a statistical model and a LULC simulation model to predict that malaria risk may be substantially higher in a governance (GOV) scenario versus a business-as-usual (BAU) scenario, a direct consequence of the averted deforestation under the GOV scenario.

In the third chapter, we discuss how current methodology to characterize mosquito breeding habitat may be improved when the goal is to guide larva control interventions and identify areas with higher disease risk. To accomplish these goals, we contend that it is critical for researchers to understand the spatial and temporal distribution of water bodies. We use simulations and *Anopheles darlingi* larva data to illustrate how inference might be misleading when the distribution of water bodies is ignored.

Finally, the fourth chapter re-examines results from a key study conducted in the Peruvian Amazon, which has been the basis for the widespread notion that malaria risk increases in deforested areas. Differently from the original studies, I integrate mosquito larva and mosquito biting rate from multiple vector species with malaria prevalence data, to determine how LULC and climate affect malaria risk. I find that *A. darlingi* larva

presence and biting rate indeed increase in deforested areas, agreeing with the original published results. However, we find that malaria prevalence is not associated with *A. darlingi* biting rate whereas it tends to be positively associated with *A. rangeli* biting rate. We hypothesize that there is a mismatch between location and time of the day at which individuals are more exposed to infections and our predictions of mosquito biting rate, which are based on household locations and a 18 – 24:00 period. If this hypothesis is correct, any association between malaria prevalence and the mosquito biting rate is likely to be spurious. Although our analysis is not conclusive because of limitations in the malaria prevalence dataset, our results suggest that inference on malaria risk based solely on a single vector species might be misleading. In particular, the relationship between deforestation and malaria risk will critically depend on how exposure changes as a function of human behavior and on the competence of the different vector species.

Throughout this thesis, a major thrust has been to employ statistical models carefully tailored to the problems at hand, with the overarching goal of generating more reliable inference. To this end, I have created parsimonious models (e.g., using multiple shrinkage priors or using a reversible jump algorithm) which properly accommodate data idiosyncrasies (e.g., zero inflation and over-dispersion) and use, when possible, data from multiple sources.

## **Dedication**

I dedicate this thesis to my beloved family here in the U.S. (Natercia, Matheus and Samuel) and in Brasil (Cleria, Luiz, Lie and Carlos). Without them and their support, this work would not have been possible.

# Contents

Abstract .....	iv
List of Tables .....	xii
List of Figures .....	xiii
Acknowledgements .....	xvi
Chapter 1. Enhanced understanding of infectious diseases by fusing multiple datasets: a case study on malaria in the western Brazilian Amazon region.....	1
1.1 Introduction.....	1
1.2 Methods .....	5
1.2.1 Data .....	5
1.2.2 Model description .....	6
1.2.2.1 Plasmodium detection .....	6
1.2.2.2 Infection risk.....	8
1.2.2.3 Symptomatic status .....	9
1.2.2.4 Likelihood.....	9
1.2.2.5 Full model.....	11
1.2.3 Model performance.....	13
1.3 Results .....	14
1.3.1 Model performance.....	14
1.3.2 Findings from the Western Brazilian Amazon region .....	18
1.4 Discussion.....	28
1.5 Citation.....	33

Chapter 2. Conservation efforts may increase malaria burden in the Brazilian Amazon	35
2.1 Introduction.....	35
2.2 Methods .....	37
2.2.1 Malaria Data.....	37
2.2.2 Catchment area.....	38
2.2.3 Covariates.....	39
2.2.4 Permutation tests.....	39
2.2.5 Regression model .....	40
2.2.6 Land use / land cover (LULC) future scenarios .....	43
2.3 Results .....	43
2.4 Discussion.....	48
2.5 Citation.....	56
Chapter 3. Abundance of water bodies is critical to guide larva control interventions and predict disease risk .....	57
Chapter 4. Revisiting the relationship between land cover and malaria in the Peruvian Amazon: the importance of integrating data on multiple vector species, mosquito life stages, and malaria prevalence .....	65
4.1 Introduction.....	65
4.2 Methods .....	66
4.2.1 Data .....	66
4.2.2 Regression models.....	68
4.2.2.1 Larva model.....	69
4.2.2.2 Mosquito biting rate model .....	69

4.2.2.3 Malaria prevalence.....	71
4.2.2.4 Covariates.....	73
4.2.2.5 Validation.....	75
4.3 Results .....	76
4.3.1 Parameter estimates and individual climate and LULC effects .....	76
4.3.2 Spatial patterns .....	82
4.3.3 Malaria prevalence and mosquito biting rate .....	84
4.4 Discussion.....	86
Appendix I .....	93
1. Description of covariates .....	93
2. Derivation of the likelihood .....	96
3. Likelihood formulae.....	97
4. Full conditional distribution for the parameters sampled via a Gibbs sampling step	99
5. Description of how data were simulated .....	99
Appendix II.....	101
1. Comparison of modeling results using different radii for the catchment area. 101	
2. Alternative model formulation.....	102
3. Auxiliary figures and tables.....	104
Appendix III.....	106
1. Simulation results.....	106
2. Description of the binomial model used to fit the <i>A. darlingi</i> larva data.....	109

2.1.	Data .....	109
2.2.	Covariates.....	109
2.3.	Methods .....	110
2.4.	Results.....	111
Appendix IV .....		114
1.	Priors for the larva model.....	114
2.	Priors for the biting rate model .....	115
3.	Latent state representation and priors for the probit regression used to model malaria prevalence .....	115
4.	Model fit.....	116
5.	Description of climate covariates .....	117
References. ....		119
Biography.....		137

## List of Tables

Table 1: List of all the estimated parameters and the associated priors.....	12
Table 2: Description of all the modeling approaches employed in the simulation and validation exercises.....	13
Table 3: Summary statistics for the estimated parameters.....	20
Table 4: Summary description of the malaria dataset. ....	38
Table 5: Summary of regression parameter estimates.....	44
Table 6: Sample of studies on mosquito breeding habitat. ....	57
Table 7: Description of outcomes for scenarios 1 and 2.....	60
Table 8: Summary of the data and model parameters.....	72
Table 9: Likelihood of each of the possible outcomes in AACD.....	98
Table 10: Likelihood of each of the possible outcomes in ACD and PCD. ....	98
Table 11: Summary of parameter values adopted for the simulated data.....	100
Table 12: Number of microscopy and PCR results for the different sampling designs, both from the original and simulated datasets.....	100
Table 13: Summary statistics of the posterior distribution of the pooled forest cover effect ( $\beta_1^{for}$ ) and deforestation rate effect ( $\beta_1^{def}$ ) for the alternative model. ....	103
Table 14: Convergence statistic R (91) for the regression parameters (intercept and slopes for the different covariates) in the main model. ....	104

## List of Figures

Figure 1: Graphical representation of the proposed model, illustrating some of the modeled conditional relationships. ....	10
Figure 2: Comparison of models using simulated data.....	16
Figure 3: Comparison of models by out-of-sample prediction. ....	17
Figure 4: Comparison of summary statistics calculated directly from the data and generated by the proposed model.....	19
Figure 5: Posterior distribution of infection risk factors.....	22
Figure 6: Posterior distribution of the other estimated parameters.....	23
Figure 7: Probability of infection $p(I = 1)$ as a function of the most important covariates. ....	24
Figure 8: Probability of sampling an asymptomatic <i>Plasmodium</i> carrier (i.e., $p(S = 0, I = 1)$ ). ....	25
Figure 9: Spatial distribution of infection, asymptomatic carrier, and malaria prevalences.....	27
Figure 10: Malaria incidence is higher in areas with more forest cover whereas no clear pattern arises regarding deforestation rates. ....	45
Figure 11: Malaria incidence tends to be higher for cities close to protected areas (PA's). ....	46
Figure 12: Predicted malaria incidence in urban health posts is higher in the governance scenario than in the business-as-usual scenario. ....	47
Figure 13: Malaria incidence increase at urban health posts in the governance scenario is predicted to be a direct consequence of prevented deforestation.....	48
Figure 14: The number of water bodies with larva per transect (lower panels) is influenced by the relationship between the proportion of water bodies with larva and	

forest area (upper panel) <i>and</i> the relationship between number of water bodies per transect and forest area (middle panels).....	62
Figure 15: Posterior distribution of slope parameters, stratified by covariate and mosquito species, for the larva model. ....	77
Figure 16: Abundance of water bodies with larva decreases with forest cover (right panel) and NFV cover (left panel). ....	78
Figure 17: Posterior distribution of slope parameters $\beta_{2_s}$ showing how mean predicted number of water bodies with larva $\hat{L}_{sji}$ affects the probability of young adults $\theta_{sji}$ .....	79
Figure 18: Posterior distribution of slope parameters $\beta_{3_s}$ showing how biting rate, given larva habitat suitability, is affected by the different covariates.....	80
Figure 19: Predicted mosquito biting rate as a function of distance to large water bodies (left panel), to forest (middle panel), and impervious areas (right panel).....	82
Figure 20: Spatial prediction <i>A. darlingi</i> in relation to the number of water bodies with larva (left panel) and log mosquito biting rate (right panel). ....	83
Figure 21: Spatial prediction of the log number of water bodies with larva (upper panels) and log mosquito biting rate (lower panels) for anopheline species.....	84
Figure 22: Predictors of malaria infection.....	85
Figure 23: Posterior distribution of the main regression parameters with covariates and population size assessed using three different catchment area radii (10, 20, and 30 km). ....	102
Figure 24: Comparison of the data (black line) and the 95% posterior predictive interval (red lines) for 20 randomly chosen cities.....	104
Figure 25: Gross domestic product is similar in cities with low and high forest cover. .	105
Figure 26: Regression models detect a significant difference between forested and deforested sites (upper panels), despite the same average per transect (lower panels).	108

Figure 27: Comparison of the predicted and observed average number of water bodies with *Anopheles darlingi* larva, for different collections periods (upper panel) and locations (lower panel)..... 112

Figure 28: Posterior distribution of slope parameters. .... 113

Figure 29: Comparison of the temporal trend on larva data (black circles) and the posterior predictive distribution (red lines)..... 116

Figure 30: Comparison of the temporal trend on mosquito biting rate data (black circles) and the posterior predictive distribution (red lines)..... 117

## **Acknowledgements**

I am grateful to those who have always been there for me when I needed them, in particular Meg Stephens, Andy Minnis, and my advisor James Clark.

# **Chapter 1. Enhanced understanding of infectious diseases by fusing multiple datasets: a case study on malaria in the western Brazilian Amazon region**

## **1.1 Introduction**

Extensive syndromic sentinel surveillance data are often routinely collected by public health agencies. However, estimates of disease prevalence based on these data are known to be biased because only symptomatic individuals are sampled (1, 2). Furthermore, because of the sentinel surveillance network extent, cheaper and less sensitive diagnostic methods are typically employed. Researchers also collect data to study infectious disease risk factors and asymptomatic pathogen carriers, but using cross-sectional surveys and more expensive and sensitive diagnostic methods. These data, however, are often geographically and temporally limited and thus are not as abundant as sentinel surveillance data. Robust inference on disease prevalence and risk factors would ideally combine these datasets because they clearly complement each other; unfortunately, standard statistical tools are not well suited for this task. We describe here a novel statistical model that coherently combines these disparate datasets, allowing for enhanced inference on infectious diseases.

Our study focuses on malaria. Malaria is responsible for ~3% of the total global disease burden (3), affecting approximately half of the world's population (4) and significantly hindering economic and social development of tropical countries (5).

Despite its public health relevance and recent increased attention to malaria research and control (6), malaria risk factors remain difficult to evaluate, due both to the idiosyncrasies of how data are collected (as detailed below) and the fact that not all infected individuals are symptomatic. Our approach addresses these challenges, providing sharper inference on *Plasmodium* infection risk factors, factors determining symptom status given infection, and overall infection and disease prevalence. We first describe the statistical model, then we compare its performance against standard logistic regression using simulated and real data, and finally we apply it to a large malaria dataset collected in the Western Brazilian Amazon.

In Brasil, malaria cases are concentrated in the Amazon region (7), resulting in substantial morbidity (8, 9). Similar to other countries (e.g., India, (10)), the malaria surveillance data from the Brazilian government consist of microscopy results from predominantly symptomatic individuals, sampled through active and passive case detection (ACD and PCD, respectively). ACD data are obtained by health agents during home visits to symptomatic individuals whereas PCD data come from health facilities, visited by individuals who believe they have malaria (11). Inherent biases in both datasets make it difficult to determine overall malaria prevalence and the factors that influence it (1, 12). Aggressive active case detection (AACD) has been proposed as an alternative surveillance technique, consisting of cross-sectional surveys where all individuals are sampled, regardless of symptom status (11). AACD data can be used to

estimate infection prevalence and its determinants and the size of the reservoir represented by asymptomatic *Plasmodium* carriers (12-14). Drawbacks of AACD include high costs and the often low acceptability from the population (12, 14), which often limits AACD data to a short time-frame and a small geographical area. As a consequence, AACD data might not be as well suited as ACD/PCD data in determining the effect of covariates that change substantially in time and/or space (e.g., precipitation and presence of wetlands).

Imperfect *Plasmodium* detection is a concern for all surveillance methods. The Brazilian Health Ministry primarily makes use of microscopy of thick blood smears, because it is relatively inexpensive and straight-forward (15). However, microscopy has limited ability to detect the pathogen when parasitemia is low (16-18). In research settings, Polymerase Chain Reaction (PCR) has been extensively used as the standard against which the sensitivity and specificity of other detection methods (e.g., microscopy and rapid diagnostic tests) are evaluated. Unfortunately, PCR data is often not available due to costs and expertise required for the procedure (19, 20).

How does one integrate the less biased but more limited dataset (e.g., data from AACD) with a more extensive, time continuous and biased dataset (e.g., data from ACD/PCD)? Furthermore, how can the more sensitive but limited PCR dataset be used jointly with the less sensitive but more extensive microscopy dataset? Logistic regression is the most common statistical tool used to analyze individual-level disease data.

However, logistic regression does not correct for the biases in the ACD/PCD dataset, even if dummy covariates are added to represent differences in how individuals were sampled (e.g., AACD, ACD, and PCD). It also does not accommodate detection error rates for the different *Plasmodium* detection methods. In recognition of these problems, analysis might focus on the most sensitive pathogen detection method (i.e., PCR) and less biased case detection method (i.e., AACD), with the drawback of ignoring considerable information contained in the rest of the data.

Logistic regression also does not allow for important conditional relationships that determine malaria risk. Malaria researchers typically assume perfect detection and choose to model either the probability of being diseased (i.e.,  $p(S = 1, I = 1)$ ) or the probability of being infected (i.e.,  $p(I = 1)$ ), where  $S$  and  $I$  stand for symptom and infection status. These probabilities are related and models can be developed to combine them in a statistically and biologically coherent way. Our model factors  $p(\text{Disease}) = p(S = 1, I = 1)$  as  $p(S = 1 | I = 1)p(I = 1)$ , allowing us to separately evaluate infection risk factors (i.e.,  $p(I = 1)$ ) from risk factors of symptoms given infection (i.e.,  $p(S = 1 | I = 1)$ ). This approach can provide inference on factors that influence the joint distribution of symptom and infection statuses. For example, we can coherently estimate the prevalence of asymptomatic carriers, namely  $p(S = 0, I = 1)$ , and the factors that influence it. The limitations of standard statistical tools prompted us to create a customized method to analyze our data.

Here, we illustrate how inference on malaria risk factors and infection/disease prevalence can be improved using a hierarchical framework based on the joint distribution of symptom and infections statuses and by properly accommodating the different pathogen and case detection methods. First, we detail the model. Then, we compare the performance of this method to that of typical logistic regressions using simulated and real data. Finally, we apply this model on a large malaria dataset collected in the Western Brazilian Amazon and discuss the implication of our findings.

## **1.2 Methods**

### **1.2.1 Data**

Data were collected in a rural settlement area, in a region known as Ramal Granada (Acrelandia, Acre, Brasil), on 486 individuals that agreed to participate in the study. AACD data come from four cross-sectional surveys (March/April 2004, September/October 2004, February/March 2005, and October/November 2006) in which all study participants that were present at the time of the survey were sampled, regardless of their symptomatic status. This dataset contained a total of 1383 microscopy and 1400 PCR malaria tests. Further details on the area, data collection, and characteristics of this cohort can be found elsewhere (11, 21, 22). We gathered ACD/PCD data by searching the malaria records at the local health facility. All malaria records between 2004 and 2007 from the AACD study participants were entered in a database,

resulting in a total of 1694 microscopy tests, with approximately 94% of the individuals feeling symptomatic when tested.

## 1.2.2 Model description

We start by describing some basic conditional probabilities for our model and their associated assumptions. We then proceed to detail the likelihood associated with each potential outcome. We conclude this section with a description of how we fit the model.

### 1.2.2.1 Plasmodium detection

We consider data from two *Plasmodium* detection methods, namely microscopy and polymerase chain reaction (PCR). Let  $D_{i,t}^m = 1$  stand for a positive *Plasmodium* detection using microscopy for individual  $i$  at time  $t$ . Let  $I_{i,t} = 1$  and  $S_{i,t} = 1$  stand for being infected and having malaria symptoms, respectively. Note that  $I_{i,t}$  is a latent variable because we never directly observe it. Using these definitions, let

$p(D_{i,t}^m = 1 | S_{i,t} = 1, I_{i,t} = 1) = \alpha_1$  and  $p(D_{i,t}^m = 1 | S_{i,t} = 0, I_{i,t} = 1) = \alpha_0$  be the microscopy sensitivity given that  $S_{i,t} = 1$  and  $S_{i,t} = 0$ , respectively. We allow sensitivity to depend on symptom status because it has been shown that low-grade infections (i.e., low density of parasites in the blood) are associated with asymptomatic cases and failure to detect them with microscopy (16-18, 23). Furthermore, let  $p(D_{i,t}^m = 0 | I_{i,t} = 0) = 1$  be the microscopy specificity. We set the specificity of the microscopy to one because it is

virtually impossible for an experienced microscopist to identify malaria pathogens on a blood sample from an uninfected patient, regardless of the symptomatic status of the patient (Ferreira, personal communication; (17)).

In relation to PCR, let  $D_{i,t}^{pcr} = 1$  stand for a positive *Plasmodium* detection using PCR for individual  $i$  at time  $t$ . Let the PCR sensitivity and specificity be denoted by  $p(D_{i,t}^{pcr} = 1 | I_{i,t} = 1) = \delta$  and  $p(D_{i,t}^{pcr} = 0 | I_{i,t} = 0) = \pi$ , respectively. Errors in amplification or contamination of the sample can produce both false-positives and false-negatives (17). From prior knowledge, we know that the sensitivity of PCR is greater than that of microscopy and that microscopy sensitivity is probably greater when the individual is symptomatic than when not symptomatic (i.e.,  $\delta > \alpha_1 > \alpha_0$ ) (20). Finally, we assume that PCR sensitivity and specificity are not influenced by microscopy detection and symptomatic status of the individual, given infection status. The assumption of conditional independence between PCR and microscopy results seems reasonable because detections are based on fundamentally different biological processes (24, 25). We adopted uniform priors for the sensitivity and specificity of PCR, where the limits were based on earlier reports on PCR error rates (26, 27). More specifically, the joint prior adopted for these detection parameters was a uniform distribution in the set  $\{(\delta, \alpha_1, \alpha_0) : 0 < \alpha_0 < \alpha_1 < \delta, \max(0.7, \alpha_1) < \delta < 1\}$ .

### 1.2.2.2 Infection risk

We are primarily interested in the probability that individual  $i$  at time  $t$  is infected with *Plasmodium* (i.e.,  $p(I_{i,t} = 1)$ ) and the associated risk factors. We assume that this probability is given by

$$p(I_{i,t} = 1) = \frac{1}{1 + e^{-(\mathbf{x}_{i,t}^T \boldsymbol{\beta} + \phi_i + \mathcal{G}_{h[i]})}}$$

where  $\mathbf{x}_{i,t}$  is the design vector and  $\boldsymbol{\beta}$  is the vector with the corresponding parameters.

The design vector  $\mathbf{x}_{i,t}$  contains potential risk factors. For our case study using data from the Western Brazilian Amazon, these covariates were gender, educational level, age, time in Acrelandia (as a proxy for past exposure to malaria), if participates on extractivism activities, if hunts or fishes, if works as chain sawyer, if shares the house with somebody that had a positive malaria diagnosis in the past 30 days, surface water area, forest area, deforestation rate, precipitation, and a drought index. These covariates are detailed in Appendix I. Individual and household-level random effects are denoted by  $\phi_i$  and  $\mathcal{G}_{h[i]}$ , respectively, where  $h[i]$  indexes the household where the  $i^{th}$  person resides. These random effects were modeled as  $\phi_i \sim N(0, \sigma_{ind}^2)$  and  $\mathcal{G}_{h[i]} \sim N(0, \sigma_h^2)$ , where  $\sigma_{ind}^2$  and  $\sigma_h^2$  are the individual and household-level random effect variances, respectively.

### 1.2.2.3 Symptomatic status

We assume that the probability of being symptomatic given that the person is infected is given by

$$p(S_{i,t} = 1 | I_{i,t} = 1) = \frac{1}{1 + e^{-\mathbf{z}_{i,t}^T \boldsymbol{\gamma}}}$$

where  $\boldsymbol{\gamma}$  is a vector of parameters to be estimated and  $\mathbf{z}_{i,t}$  is the design vector. We assume that the covariates most likely to influence this probability are variables related to the individual's immune system and not variables related to present exposure to vectors. Thus, for our Western Brazilian Amazon case study, the covariates in  $\mathbf{z}_{i,t}$  were age, gender, and time in Acrelandia (as a proxy for past malaria exposure). Finally, we assumed that the probability of having symptoms despite not being infected

$p(S_{i,t} = 1 | I_{i,t} = 0)$  was a constant parameter to be estimated.

### 1.2.2.4 Likelihood

The definitions above are the basis for the hierarchical model that we built (depicted in Figure 1), borrowing some ideas from Clark & Hersh (28). These definitions and model structure allow us to describe the likelihoods of all the possible outcomes in AACD (Table 9 in Appendix I). For the ACD and PCD datasets, we start by noting that  $p(I | ACD) > p(I)$  and  $p(I | PCD) > p(I)$ , because ACD and PCD focuses mostly on symptomatic individuals. Therefore, we can assume that knowing whether the person was sampled in ACD or PCD does not bring any additional information about the risk of

being infected if we condition on symptomatic status. More formally, we assume that  $p(I | S, ACD) = p(I | S)$  and  $p(I | S, PCD) = p(I | S)$ . Based on these assumptions, it can be shown that the likelihood for each outcome will be similar to those for AACD

with the exception that it will have a correction term of the form  $\frac{p(ACD | S)}{p(ACD)}$  or

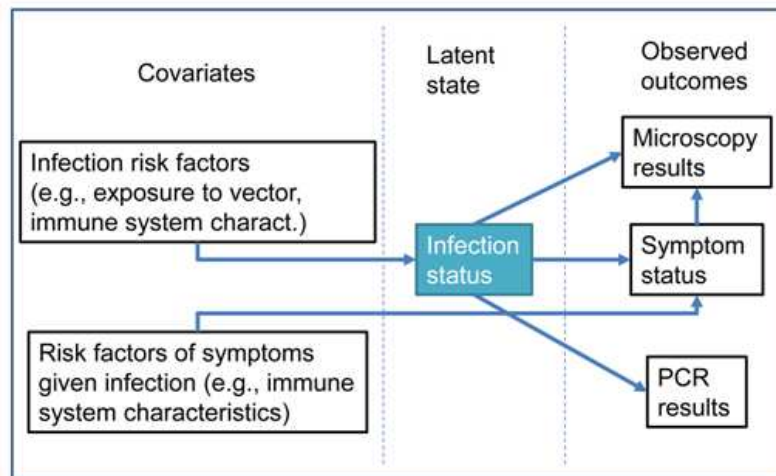
$\frac{p(PCD | S)}{p(PCD)}$ . Here,  $p(ACD | S)$  and  $p(PCD | S)$  are the conditional probability that an

individual with symptom status  $S$  is sampled through ACD or PCD, respectively, and

$p(ACD)$  and  $p(PCD)$  are the corresponding marginal probabilities. The likelihood of

all the possible outcomes in ACD and PCD is shown in Table 10 in Appendix I. The

detailed derivation of these likelihoods is also given in Appendix I.



**Figure 1: Graphical representation of the proposed model, illustrating some of the modeled conditional relationships.**

An important assumption in our analysis of the ACD/PCD dataset is that malaria tests (and the symptomatic status at the time of the test) more than one week apart from each other were considered to be independent. There were some cases where symptomatic individuals would choose to be tested multiple times within a short period of time ( $< 7$  days), probably expecting a positive result or the symptoms to ameliorate. To avoid making several assumptions regarding the temporal dependencies of symptoms and test results from these multiple tests, we chose to retain just the first test and the associated symptomatic status whenever we detected multiple tests within this short time-frame.

#### 1.2.2.5 Full model

Let  $\theta$  be all the parameters we will estimate and let  $y_{PCD}$ ,  $y_{ACD}$  and  $y_{AACD}$  be the different datasets, where subscripts denote how individuals were sampled. Assuming conditional independence given the parameters  $\theta$ , the full model can be written as

$$\begin{aligned} p(\theta | y_{ACD}, y_{PCD}, y_{AACD}) &\propto p(y_{ACD}, y_{PCD}, y_{AACD} | \theta) p(\theta) \\ &\propto p(y_{ACD} | \theta) p(y_{PCD} | \theta) p(y_{AACD} | \theta) p(\theta) \end{aligned}$$

where  $p(\theta | y_{PCD}, y_{ACD}, y_{AACD})$  is the posterior distribution of the parameters to be estimated,  $p(y_k | \theta)$  is the likelihood of dataset  $k$  (Table 9 and Table 10 in Appendix I) and  $p(\theta)$  are the priors. All the estimated parameters  $\theta$  are listed and described in Table 1, together with their associated prior distributions.

**Table 1: List of all the estimated parameters and the associated priors.**

Parameter	Description	Prior
$\alpha_1$	Microscopy sensitivity given S=1	uniform in the set $\{(\alpha_0, \alpha_1, \delta) : 0 < \alpha_0 < \alpha_1 < \delta, \max(0.7, \alpha_1) < \delta < 1\}$
$\alpha_0$	Microscopy sensitivity given S=0	
$\delta$	PCR sensitivity	
$\pi$	PCR specificity	Unif(0.97,1)
$\beta$	Covariates of infection risk factors	Unif(-10,10)
$\varphi_i$	Individual level random effects	$N(0, \sigma_{ind}^2)$
$\mathcal{G}_{h[i]}$	Household level random effects	$N(0, \sigma_h^2)$
$\sigma_{ind}$	Standard deviation of the individual-level random effects	Unif(0,100)
$\sigma_h$	Standard deviation of the household-level random effects	Unif(0,100)
$\gamma$	Covariates of risk factors of symptoms given infection	Unif(-10,10)
$p(S=1   I=0)$	Probability of symptoms given no infection	Unif(0,1)
$p(PCD   S=0)$	Probability of being sampled through PCD given no symptoms	uniform in the set $\{(p(PCD   S=0), p(ACD   S=0)) : p(ACD   S=0) + p(PCD   S=0) < 1\}$
$p(ACD   S=0)$	Probability of being sampled through ACD given no symptoms	
$p(PCD   S=1)$	Probability of being sampled through PCD given symptoms	uniform in the set $\{(p(PCD   S=1), p(ACD   S=1)) : p(ACD   S=1) + p(PCD   S=1) < 1\}$
$p(ACD   S=1)$	Probability of being sampled through ACD given symptoms	

This model was fitted using a Gibbs sampler. Most parameters were updated using a Metropolis sampling step and the few parameters that were updated via a Gibbs sampling step have their full conditional distributions described in Appendix I. In total, 150,000 iterations were run and the initial 20,000 iterations were discarded as burn-in. Convergence was assessed using trace-plots of the parameters.

### 1.2.3 Model performance

We compare the proposed model with standard logistic regressions, both with and without individual and household level random effects. Let  $D = 1$  be a positive *Plasmodium* detection, either from microscopy, PCR or both. The response variable for these logistic regressions were proxies for a) disease: a person having symptoms and a positive detection (i.e.,  $D = 1, S = 1$ ); and b) infection: a person having a positive detection (i.e.,  $D = 1$ ) (Table 2). To mimic how researchers would typically use these multiple datasets ( $y_{ACD}$ ,  $y_{PCD}$  and  $y_{AACD}$ ), we merged the three datasets into a single one and added two dummy covariates in the logistic regressions to allow for differences between datasets.

**Table 2: Description of all the modeling approaches employed in the simulation and validation exercises.**

† these models were fit using the ‘glm’ function in R; †† these models were fit using the ‘lmer’ function in R.

Models	Outcome	Description	Random effects
1	$(D = 1, S = 1), (D = 1, S = 0), (D = 0, S = 1), (D = 0, S = 0)$	proposed model	Yes
2†	Disease ( $D = 1, S = 1$ )	logistic regression	No
3†	Infection ( $D = 1$ )	logistic regression	No
4††	Disease ( $D = 1, S = 1$ )	logistic regression	Yes
5††	Infection ( $D = 1$ )	logistic regression	Yes
No covariate	Disease ( $D = 1, S = 1$ ) Infection ( $D = 1$ )	Uses the proportion of $(D = 1, S = 1)$ and $(D = 1)$ in the training dataset to predict outcomes for the validation dataset	No

These different statistical methodologies were compared using both simulated and real data. Simulated data were used to compare the different methods in relation to

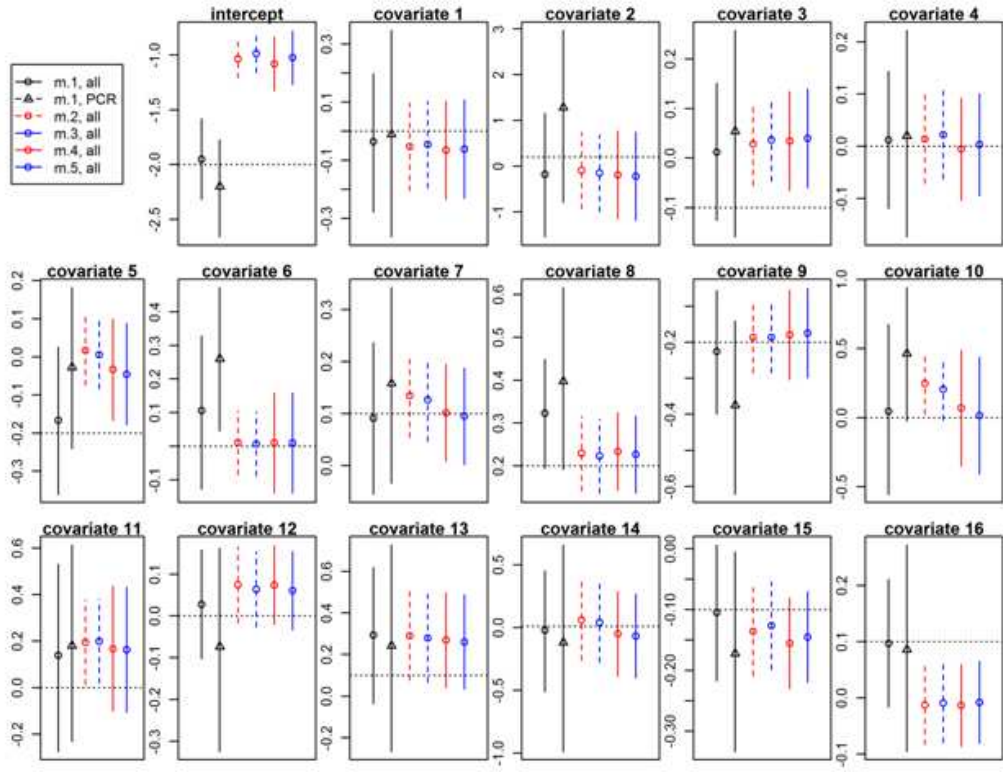
how well they retrieved the true parameters influencing infection probability. To evaluate the importance of combining these multiple datasets, we further compared how inference from the proposed model would change if fitted only to the PCR dataset versus all datasets. Details of how the simulated data were generated are given in Table 11 and Table 12 within Appendix I. We also compared how well each model predicted the real data, using a 10-fold cross validation. This validation exercise consisted in fitting these models to 90% of the real data and comparing their predictions for the remaining 10%. This was done ten times with different portions of the data retained for validation at each time. Each method predicted which individuals had a positive test result ( $D = 1$ ) and which individuals had a positive test result and were symptomatic ( $D = 1, S = 1$ ). We summarized this information as a) the proportion of individuals correctly predicted as  $D = 1$  or  $D = 0$ ; and b) the proportion of individuals correctly predicted as  $D = 1, S = 1$  or not  $D = 1, S = 1$ . For this validation exercise, we also evaluated the predictive ability of the chosen covariates by adding the prediction results from a model that simply used the proportion of individuals with  $D = 1$  (or  $D = 1, S = 1$ ) in the training dataset. All statistical procedures and graphics were performed in R (29).

## **1.3 Results**

### **1.3.1 Model performance**

Our results using simulated data reveal that the 95% confidence intervals from the logistic regressions, both with and without random effects, were typically narrower

than the 95% credible intervals from the proposed model (Figure 2), often missing the true regression parameters, even when these effects were large. In contrast to these results, the 95% credible interval generated by the proposed model fitted to all datasets always included the true regression parameters. One parameter of particular importance is the intercept as it reveals the infection prevalence for individuals with mean covariate values. Our results show that all logistic regressions grossly overestimated this parameter. The simulated data also revealed that fitting the proposed model to all datasets (microscopy and PCR results from the ACD, PCD, and AACD datasets) resulted in sharper inference, both in terms of smaller bias and uncertainty, when compared to results from the proposed model fitted just to PCR results (black circle vs. black triangle, Figure 2). This improved inference arises not only because of the larger sample size but also because the ACD and PCD datasets are more time continuous, resulting in greater variability for several covariates.

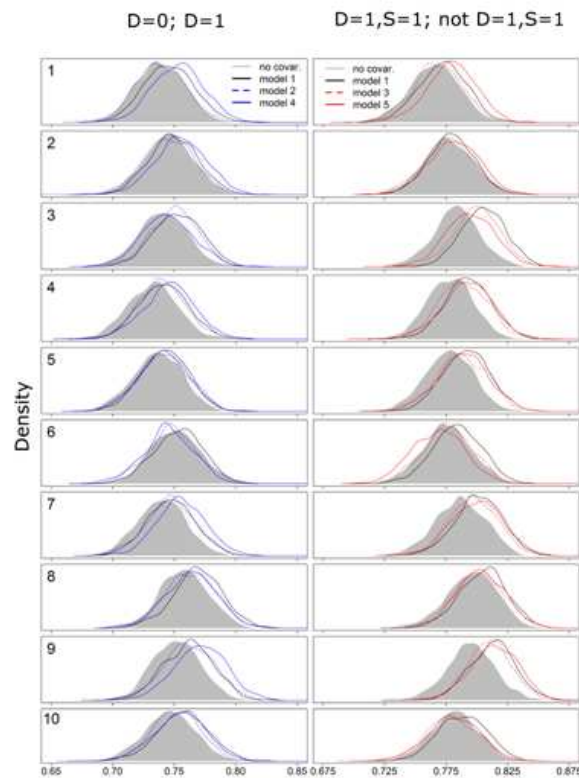


**Figure 2: Comparison of models using simulated data.**

The true values of the infection risk factor parameters are depicted in horizontal black dashed lines. Logistic regression models with disease (models 2 and 4) and infection (models 3 and 5) as response variables are depicted in red and blue, respectively. Models with and without random effects are depicted with continuous and dashed vertical lines, respectively. Models 2-5 were fitted to all datasets. Model 1 was fitted twice, once for just the PCR dataset (black triangle) and once for all datasets (black circle). Details of these models are given in Table 2.

An important concern related to the proposed model is that it might be over-fitting the data, given that it includes almost twice as many parameters as the logistic regressions (30 vs. 17, respectively, after excluding random effects and their variances), potentially resulting in poor out-of-sample predictive ability. However, our validation results using the real data show that the proposed model had a similar or better predictive ability when compared to the logistic regression model with random effects

(Figure 3). Interestingly, even the model without any covariates had a good predictive ability, sometimes yielding equivalent or better predictions than the logistic models, with or without random effects. In contrast, the proposed model always yielded better predictions than the model without any covariates. Furthermore, the proposed model is capable of generating all predictions depicted in Figure 3 whereas distinct logistic regressions were fit to predict these different outcomes.



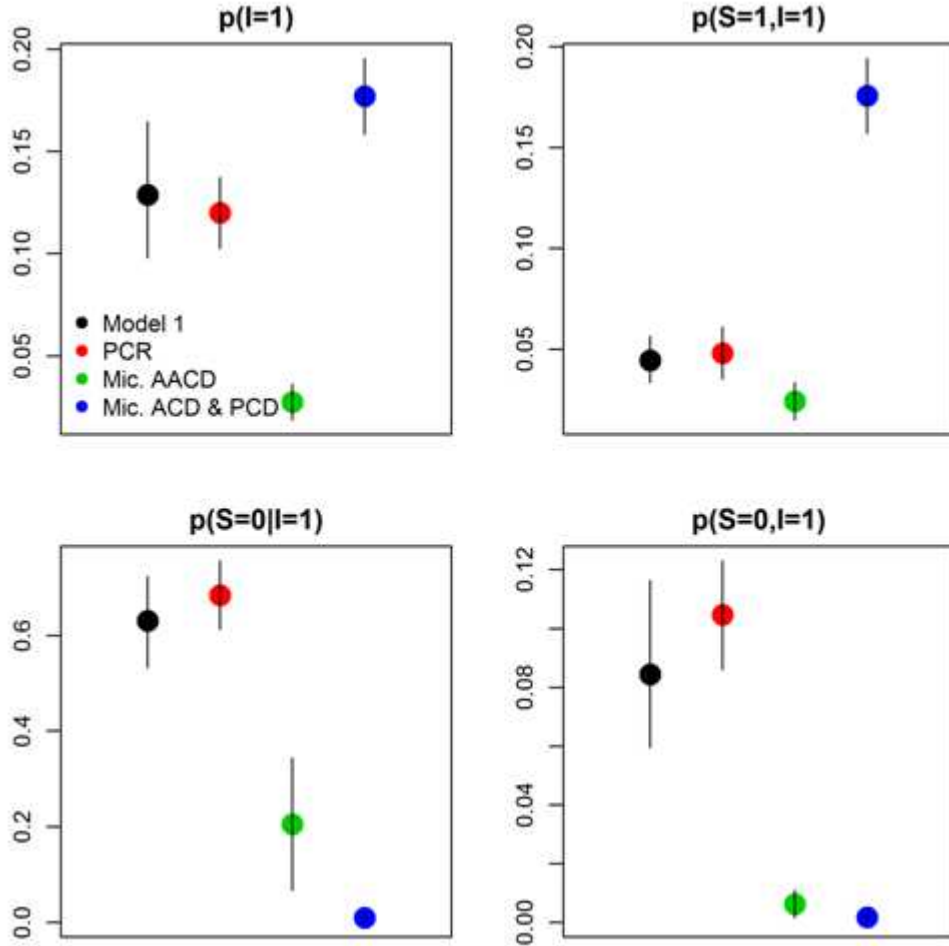
**Figure 3: Comparison of models by out-of-sample prediction.**

These figures show the proportion of individuals correctly classified by each model. Numbers on the left refer to the different validation datasets. Logistic regression models with disease (models 2 and 4) and infection (models 3 and 5) as response variables are depicted in red and blue, respectively. Models with and without random effects are depicted with continuous and dashed lines, respectively. Details of these models are given in Table 2.

### 1.3.2 Findings from the Western Brazilian Amazon region

We estimated that the infection prevalence for the cohort we studied was approximately 0.13 (95% credible interval (CI) 0.10-0.16). Malaria prevalence was considerably lower (0.04, 95% CI 0.03-0.06) because not all individuals exhibit symptoms. From the pool of infected individuals, more than half will typically be asymptomatic (0.63, 95% CI 0.53-0.72) but the overall prevalence of asymptomatic carriers is low (0.08, 95% CI 0.06-0.12). We can compare these model-based estimates with estimates calculated directly from the data, if we assume that all individuals with a positive (or negative) detection result are infected (or not infected). Similar, but not identical, results were obtained using only PCR data (Figure 4). On the other hand, considerably different summary statistics were obtained using microscopy, either from AACD or from the PCD/ACD datasets. These differences arise because microscopy is known to have limited ability to detect individuals with low parasitemia, which tend to be asymptomatic individuals, and because the PCD/ACD datasets include predominantly symptomatic individuals. One option would be to analyze just the PCR dataset collected with the AACD method, ignoring malaria risk information from the other datasets. However, as we showed with the simulated data and as suggested elsewhere (30), inference can be greatly improved when all datasets are jointly used if the model is able to adequately accommodate the inherent differences among datasets.

Thus, we exploit the information on infection/disease prevalence and malaria risk factors from all datasets.



**Figure 4: Comparison of summary statistics calculated directly from the data and generated by the proposed model.**

The summary statistics are infection (i.e.,  $p(I = 1)$ ) and malaria prevalence (i.e.,  $p(S = 1, I = 1)$ ), proportion of asymptomatic individuals among the pool of infected individuals (i.e.,  $p(S = 0 | I = 1)$ ) and overall proportion of asymptomatic carriers in the population (i.e.,  $p(S = 0, I = 1)$ ). Estimates from the proposed model are depicted in black. Estimates calculated directly from the data are depicted in red (PCR data), green (microscopy results from AACD), and blue (microscopy results from ACD and PCD). Vertical lines depict 95% credible intervals for model 1 and

approximate 95% confidence intervals for the other estimates, calculated as  $\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ .

The clearest infection risk factor was forest extent surrounding the subject's house (Table 3, Figure 7; the marginal posterior distributions for all the estimated parameters are provided in Figure 5 and Figure 6). The effect of forest extent was further exacerbated by proximity to larger water bodies, particularly during the wet season. Furthermore, men (probably as a result of spending more time in the forest than women) and those participating in forest related activities (e.g., extractivism, hunting or fishing) were more likely to be infected (Table 3, Figure 7). These risk factors consistently suggest that these degraded forests are prime breeding habitat for the vector. On the other hand, annual deforestation rates and working as a chain sawyer were not important risk factors. We hypothesize that the extensive use of fire for land clearing during the dry season might be responsible for this pattern. We also expected increased infection risk if the person co-inhabited a house with somebody diagnosed with malaria within the past 30 days but this was not the case, probably because infectious individuals might be diagnosed after (instead of before) the focal person is tested for malaria. Unfortunately, these past and future dependencies cannot currently be included in the model.

**Table 3: Summary statistics for the estimated parameters.**

Class	Parameter	Percentile		
		2.50%	50%	97.50%
Infection risk factors (odds-ratio)	Intercept	0.076	0.120	0.187
	Gender	0.430	0.657	0.986
	Age	0.878	1.096	1.345
	Education	0.841	1.008	1.215
	Time in Acrelandia	0.596	0.763	0.956
	Chain Sawyer	0.009	0.363	3.235
	Extractivism	1.057	1.782	2.994
	Hunting/Fishing	1.140	1.647	2.386
	Co-inhabits D <sup>m</sup> =1	0.559	0.917	1.490
	Co-inhabits D <sup>cc</sup> =1	0.403	0.824	1.691
	Water area	0.667	0.800	0.957
	Forest area	1.430	1.923	2.569
	Annual defor.	0.719	0.909	1.139
	Monthly precip.	0.810	0.975	1.175
	Drought index	0.770	0.932	1.145
	Precip. x forest	1.004	1.183	1.405
Drought x forest	0.803	0.958	1.155	
Water x forest	1.048	1.404	1.899	
Symptoms given infection risk factors (odds-ratio)	Intercept	0.411	0.641	1.076
	Age	0.645	0.884	1.240
	Gender	0.451	0.859	1.704
	Time in Acrelandia	1.043	1.481	2.268
Other parameters (probabilities)	Mic. sensit.   S=0	0.053	0.101	0.175
	Mic. sensit.   S=1	0.249	0.293	0.348
	PCR Sensitivity	0.708	0.796	0.901
	PCR Specificity	0.970	0.974	0.990
	p(S=1   I=0)	0.015	0.023	0.034
	p(ACD   S=1)	0.075	0.380	0.770
	p(ACD   S=0)	0.000	0.002	0.005
	p(PCD   S=1)	0.084	0.391	0.775
p(PCD   S=0)	0.000	0.001	0.002	

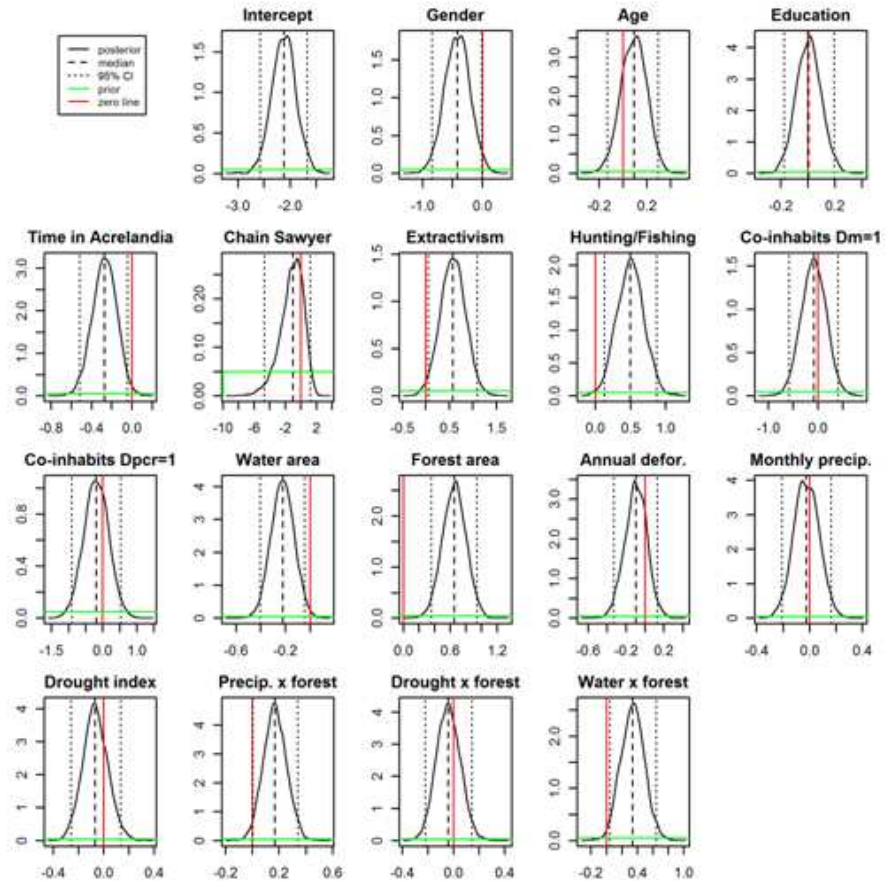


Figure 5: Posterior distribution of infection risk factors.

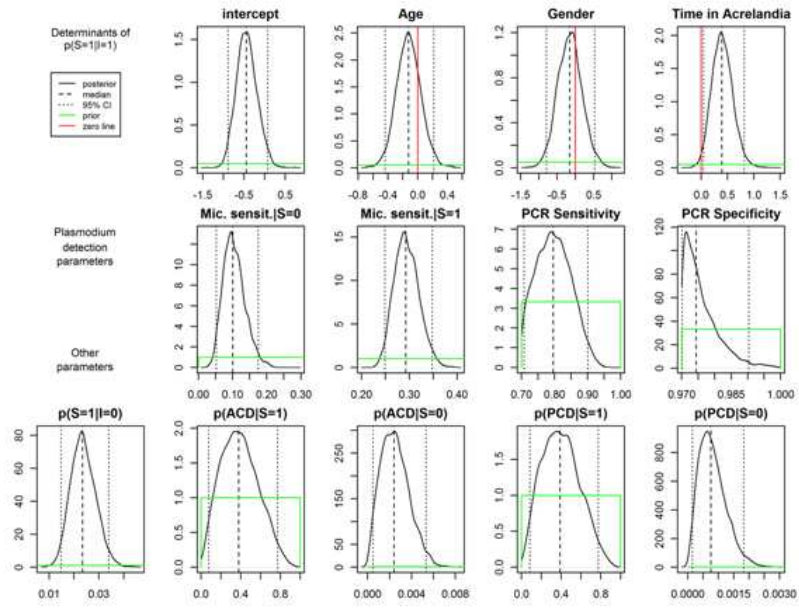
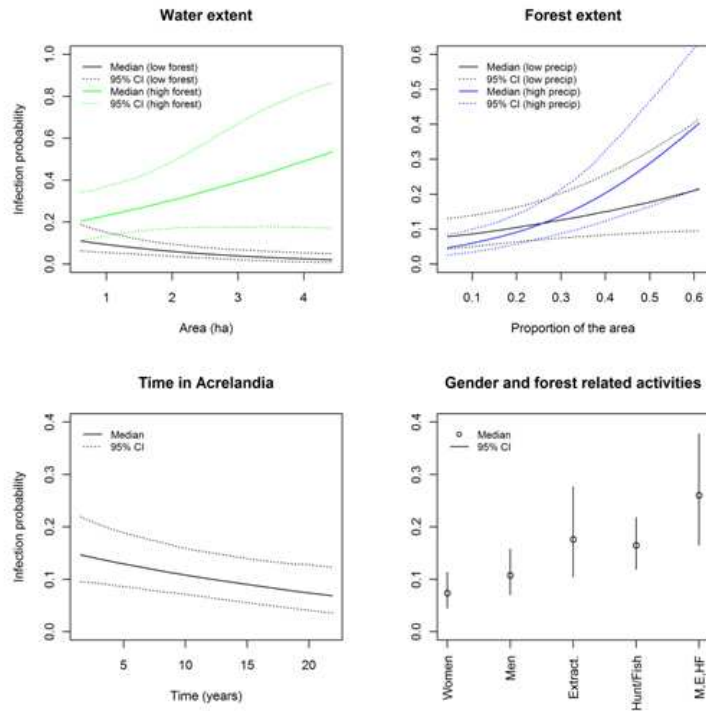


Figure 6: Posterior distribution of the other estimated parameters.

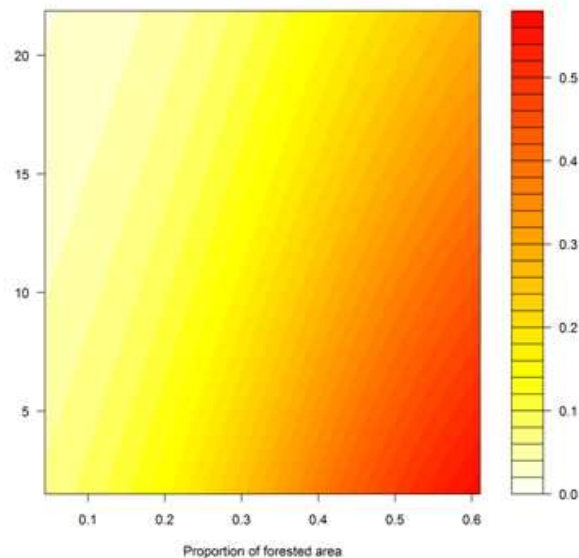


**Figure 7: Probability of infection  $p(I = 1)$  as a function of the most important covariates.**

The probability of infection was calculated with the other covariates fixed at their mean value. CI stands for credible interval. Lower right panel shows the independent effect of being a woman ('Women'), being a man ('Men'), participating on extractivism activities ('Extract.'), and participating on hunting or fishing activities ('Hunt/Fish'). The summed effect of being a man, participating on extractivism and hunting/fishing activities is also shown ('M,E,HF').

There is some evidence that time living in Acrelandia, as a proxy for past malaria exposure, reduces the risk of being infected (Table 3, Figure 7). This result suggests that non-naïve settlers acquire parasitological immunity and/or considerable knowledge on how to reduce one's exposure to infection. However, our results also suggest that this same factor increases the probability of feeling symptoms once infected (Table 3). One possible explanation is that non-naïve settlers are only susceptible to the more virulent *Plasmodium* strains.

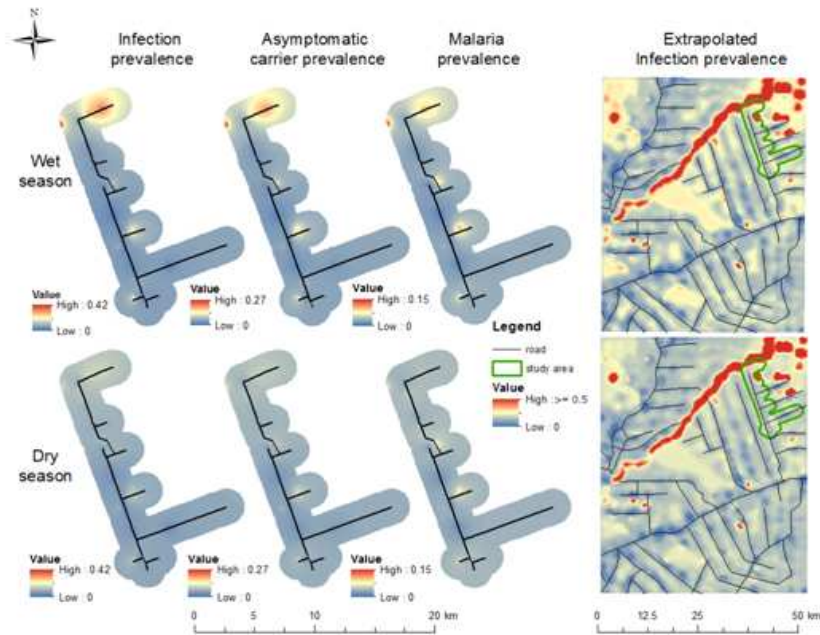
Asymptomatic *Plasmodium* carriers pose a considerable public health challenge. Our results suggest ways to strategically identify these carriers. While sampling all individuals regardless of symptoms (as in AACD) might be useful, a more efficient strategy would be to sample individuals at high risk of infection but low probability of feeling symptoms given infection. In other words, we maximize  $p(S = 0, I = 1) = p(S = 0 | I = 1)p(I = 1)$  by maximizing the individual components  $p(S = 0 | I = 1)$  and  $p(I = 1)$ . For instance, if we estimate the probability of being an asymptomatic *Plasmodium* carrier as a function of time in Acrelandia and forest extent, it becomes clear that we should preferentially sample individuals that are new to the area (thus with high  $p(S = 0 | I = 1)$ ) on highly forested areas with abundant surface water (thus with high  $p(I = 1)$ ) (Figure 8).



**Figure 8: Probability of sampling an asymptomatic *Plasmodium* carrier (i.e.,  $p(S = 0, I = 1)$ ).**

The probability of sampling an asymptomatic *Plasmodium* carrier is shown as a function of time in Acrelandia and proportion of forest area in places with abundant surface water.

The estimated parameters can be jointly used to make coherent predictions, relying on information from all datasets. For example, a predicted infection risk surface can be created using information on surface water and forest area (infection prevalence map in Figure 9). These results can be extrapolated to a larger geographical region using remote sensing imagery, revealing substantial spatial heterogeneity in infection prevalence attributable to the river that crosses the upper part of the region and the large forest blocks away from the roads (extrapolated infection prevalence map in Figure 9). These maps also highlight the striking differences in infection prevalence due to precipitation, a result greatly corroborated by recent entomological surveys conducted at the same site (31). Besides infection risk surfaces, asymptomatic carrier risk and malaria burden surfaces can also be created, using household information on how long people have been living in Acrelandia (asymptomatic carrier and malaria prevalence maps in Figure 9). Despite similarities, the asymptomatic carrier prevalence surface indicates that these carriers are more likely to be found in the northern part of our study area whereas infected symptomatic individuals can also be found in the central region of our study area.



**Figure 9: Spatial distribution of infection, asymptomatic carrier, and malaria prevalences.**

From left to right, maps depict interpolated surfaces of predicted infection prevalence (i.e.,  $p(I = 1)$ ), asymptomatic carrier prevalence (i.e.,  $p(S = 0, I = 1)$ ), and malaria prevalence (i.e.,  $p(S = 1, I = 1)$ ), all for the studied area, and an extrapolated surface of infection prevalence. Upper and lower maps are the prevalence surfaces for the rainy and dry seasons, respectively. Interpolation was done using an inverse distance weighted algorithm.

As expected, we find strong influence of priors on the estimation of the PCR error rates (Figure 6), suggesting that there was not enough information on our dataset to estimate all these parameters jointly. Microscopy sensitivity, on the other hand, was well estimated to be approximately 0.3 and 0.1, almost a three-fold difference for symptomatic and asymptomatic individuals, respectively (Table 3). Nevertheless, even for symptomatic individuals, sensitivity of microscopy was relatively low. Several quantities can be derived from these error rate estimates. For example, sampling

predominantly symptomatic individuals (as is usually done in ACD/PCD) is sensible given that the probability of being infected  $p(I = 1) = 0.13$  increases dramatically if the person is symptomatic  $p(I = 1 | S = 1) = 0.63$ . However, the challenge of using microscopy as the only method to monitor infection and disease prevalence is evident if we compare our knowledge of infection probability for symptomatic individuals before and after obtaining a negative microscopy result ( $p(I = 1 | S = 1) = 0.63$  and  $p(I = 1 | S = 1, D^m = 0) = 0.55$ , respectively), indicating very little gain of information when microscopy yields a negative result. This finding suggests that close monitoring of individuals that are symptomatic but that have recently obtained a negative microscopy result might be warranted. On the other hand, a positive microscopy detection is very informative since  $p(I = 1 | D^m = 1) = 1$ . PCR results, regardless if positive or negative, were also informative since  $p(I = 1) = 0.13$  but  $p(I = 1 | D^{pcr} = 1) = 0.76$  and  $p(I = 1 | D^{pcr} = 0) = 0.03$ .

## **1.4 Discussion**

Large spatial-scale patterns regarding malaria typically involves syndromic surveillance data (e.g., (32, 33)), despite limited microscopy sensitivity and the biased nature of these data. On the other hand, more reliable infection and disease prevalence estimates are often spatially and temporally restricted, relying almost exclusively on PCR data (13, 22, 34-36). The proposed model uses information from both datasets to improve the estimates of infection and disease prevalence at our research site, which is

then extrapolated to a larger area. Alternatively, we can infer large spatial-scale patterns of *Plasmodium* infection prevalence using the syndromic surveillance data *after* adjusting for the inherent biases in this dataset. This adjustment is only possible with the parameters estimated here and is part of our ongoing research.

Our results identify the important role of forests and forest related activities in *Plasmodium* infection risk, particularly during the rainy season and in close proximity to large water bodies (Figure 7). Unfortunately, the data do not contain more information regarding these forests (e.g., level of forest degradation) and thus we cannot determine which characteristic of these forests are important infection risk factors. These results corroborate the findings of others that proximity to the forest enhances infection risk (22, 31, 37-39) but we do not find support for the idea that deforestation activity *per se* (33) or the lack of forest (40, 41) significantly increase infection risk. Our results also suggest that one of the factors most amenable to public policy is the participation in forest related activities (e.g., extractivism, hunting and fishing activities). Hunting and fishing activities are particularly popular, with nearly two thirds of the individuals in our cohort reporting that they engage in these activities. Educational campaigns might be effective in raising awareness about how participation in these activities affects one's health and the health of their family and community, particularly for those individuals more likely to exhibit symptoms given infection (i.e., non-naïve settlers).

Malaria immunity is typically portrayed as a phenomenon that depends on age (as a proxy for past malaria exposure), with severe malaria being relatively common for young children, and older cohorts having progressively less cases of severe malaria and proportionally more cases of mild malaria and asymptomatic infection (23, 42). This descriptions refers to people exposed to malaria since birth in holoendemic countries, but it is much more complex (and less well understood) in areas with lower levels of exposure and where mild malaria predominates (23). In these latter settings, previous studies have suggested that past exposure to malaria can decrease clinical malaria risk in rural settlers (22) and provide both anti-parasite and anti-disease immunity in traditional riverine populations (13, 34). Our results suggest that anti-parasite immunity arises even in rural settlers. However, unlike previous studies, we find evidence that it also increases the probability of feeling symptoms once infected. We hypothesize that more experienced settlers are susceptible only to more virulent *Plasmodium* strains. Further studies are clearly needed to determine if this hypothesis is correct.

Joint models or analyzes, like ours, are models that simultaneously make inference on multiple outcomes (e.g., detection and symptom status), even allowing one outcome to influence the others (e.g., symptom status affecting detection). These models have recently become very popular in the medical statistics literature because more information and interpretability can be gained when compared to performing separate analysis of the different outcomes (e.g., (43, 44)). Another active research area in

statistics focuses on the use of multiple pathogen detection methods to determine overall disease prevalence and sensitivity/specificity of these detection methods (24, 45-49). A recent malaria-specific example can be found in Speybroeck et al. (50). Our model builds on both of these trends by evaluating the risk factors of infection and symptoms given infection using data from multiple case and pathogen detection methods. Our results using simulated and real data revealed that the proposed model yields better inference on risk factors and disease/infection prevalence without over-fitting the data. To our knowledge, most of the epidemiological research regarding malaria has focused on infection risk factors. However, unlike standard logistic regression, the proposed model allows coherent inference on several other important parameters, such as detection error rates and risk factors associated with symptoms given infection. The latter is critical to advance our understanding of malaria burden and asymptomatic carriers. A direct result of this coherent inference is the identification of the need for better monitoring strategies regarding symptomatic individuals with negative microscopy results and how to sample more effectively potential asymptomatic *Plasmodium* carriers (Figure 8). Finally, predicted surfaces of infection risk, asymptomatic carrier risk, and malaria burden allow for optimal spatial allocation of resources and malaria control activities.

One of the critical assumptions in our analysis was that data from ACD/PCD only differ from the AACD data by the unusually high proportion of symptomatic individuals. Although this is clearly a key factor, other characteristics of the individuals

sampled in ACD/PCD might also be important, such as the distance of their house to the health facility. Also, our model clearly depends on having individual level data on both positive and negative microscopy tests. Unfortunately, individual level data from negative microscopy tests are typically discarded, both by the Brazilian Ministry of Health and malaria researchers, hampering future analysis of these rich datasets.

We modeled symptomatic status as a binary variable despite the fact that there is considerable variation in the type and intensity of symptoms one may exhibit (11). Future work might allow for multinomial or continuous symptomatic status. Evidently, this would only be productive if this symptomatic status score was collected routinely in AACD and ACD/PCD. Another variable not included in the model is parasitemia. Precise and accurate estimates of this variable can be challenging to obtain (51). Although new quantitative PCR methods can potentially overcome this problem, dramatic fluctuations in parasite density occur in the same individual within a short time period (18). Therefore, the inclusion of parasitemia into an analysis like ours remains an important challenge. Furthermore, there is no way to distinguish new *Plasmodium* infections from recrudescence and relapses, even using modern genotyping technology, given that an individual might be initially infected by multiple strains and/or re-infected by the same common strain (52, 53). Thus, what we have called infection risk factors actually refers to the risk factors of having a relapse, a recrudescence, and/or a new infection. Finally, because *P. vivax* and *P. falciparum* are

particularly prevalent in the region, it would be interesting to evaluate if the probability of feeling symptoms given infection or the infection risk factors differ among these species. This remains an important research topic.

Using malaria in the Western Brazilian Amazon as a case study, we have shown that the modeling framework presented here can exploit information from multiple datasets to shed light on several aspects of an infectious disease (e.g., infection risk factors, risk factors associated with symptoms given infection, detection error rates) that are critical for its monitoring and control (e.g., indicating how to efficiently search for asymptomatic carriers and which symptomatic individuals should be closely monitored). While standard logistic regressions are undoubtedly important tools, these statistical models are not well suited to integrate multiple datasets. We believe that the Bayesian modeling framework described here fundamentally enhances our ability to overcome this challenge, being broadly applicable to other settings and diseases whenever asymptomatic carriers are an important public health concern and multiple datasets are available.

### **1.5 Citation**

This article has been published and should be cited as “Valle, D.; Clark, J.; Kaiguang, Z. 2011. Enhanced understanding of infectious diseases by fusing multiple

datasets: a case study on malaria in the Western Brazilian Amazon region. PLOS One, 6(11): e27462''

## **Chapter 2. Conservation efforts may increase malaria burden in the Brazilian Amazon**

### ***2.1 Introduction***

Deforestation has been a major concern in much of the tropics because of its detrimental effect on biodiversity, atmospheric carbon emissions, regional weather patterns, among other ecosystem services (54-56). The Brazilian Amazon in particular has received considerable attention because a large fraction of tropical forest clearing has occurred within this region (57). This fact has prompted the creation of the world's largest forest-conservation initiative to reduce emissions from deforestation and degradation (REDD+), with an initial pledge of up to \$1 billion USD (58), and a commitment by the Brazilian government to reduce Amazon deforestation by 80% (59). However, few conservation scientists seem to be aware that the Brazilian Amazon also plays an important role in terms of malaria cases and fatalities; almost half of the deaths attributed to this disease in the Americas occurred in Brazil (60, 61) and virtually all malaria cases in Brazil originate from the Brazilian Amazon (7, 62). To reduce malaria morbidity and mortality in the region, multi-million dollar initiatives focused on malaria have also been created (e.g., \$5 million USD/year from the Amazon Malaria Initiative (63); and ~\$23 million USD from the Global Fund to Fight Aids, Tuberculosis, and Malaria (64)).

While it is generally agreed that environmental factors play an important role in malaria (3), there are mixed evidence regarding how land cover and deforestation affect malaria in the Amazon region. For instance, proximity to forest fringes (8, 31, 65-67) and land clearing (8, 22, 33, 38, 40, 41, 68-70) have both been proposed to explain malaria vector presence, mosquito biting rate and malaria incidence. Yet, the exact role of these factors on malaria incidence has important implications regarding land use land cover (LULC) policies. Based on the evidence of higher malaria risk at recently deforested areas or in areas with active land clearing, it has been suggested that forest conservation can decrease disease burden (71-76). Based on evidence of higher malaria risk when close to forest fringes, the opposite conclusion has been reached; it has been suggested that the long-term effect of land clearing is to increase the distance of humans to forest edges and thus decrease malaria risk (8, 77, 78).

These contrasting effects of deforestation have not been studied on a large spatial scale. Here we assess the magnitude of both of these malaria risk factors with a large malaria dataset (totaling ~1.3 million positive malaria tests, gathered over ~4.5 years and over a 4.5 million km<sup>2</sup> region) and evaluate the public health consequences of current and future land use land cover (LULC) scenarios.

## **2.2 Methods**

### **2.2.1 Malaria Data**

The malaria data were collected from January 2004 to August 2008 by the Brazilian malaria surveillance system (79) and are aggregated by month and health facility. A malaria case is defined as an individual that has fever and that has a positive *Plasmodium* spp. detection through microscopy (80). To the best of our knowledge, this definition has been consistently used throughout the entire 2004-2008 period. Because there are no data on the exact location of each health facility, our approach was to subset the health facilities that are known to be in the urban area and use the spatial coordinates of the corresponding cities as proxies for their location. Determining the approximate location of these health facilities is important to adequately characterize the environmental risk factors to which individuals treated at these health facilities are exposed. We emphasize that despite being classified as urban areas, these are predominantly small cities (i.e., median population size equal to 14,000 people), often surrounded by a considerable area of forest (i.e., 22% of these cities had >50% of their catchment area covered by forests). The surrounding vegetation is critical because it is common for individuals to get infected in the surrounding area (e.g., while participating on selective timber logging, non-timber forest products collection, slash-and-burn agriculture, night fishing, hunting, mining, etc.) but to be diagnosed in the city (9, 81). We further excluded cities that had less than two years of data because it would not be

possible to estimate yearly and monthly city-specific random effects for these cities. The final dataset contained approximately half of the original malaria cases (~1,300,000 cases) but covered a similar geographical area (96% of the counties in the original dataset) (a summary description of these data is available in Table 4).

**Table 4: Summary description of the malaria dataset.**

\* CA: catchment area, defined as a 20 km buffer around each city

All columns, except for the states and number of cities, are averages of quantities assessed within each catchment area at multiple time periods

States	Number of Cities	Population (in CA* per city)	Malaria cases (in CA per month per city)	Forest cover (% of CA per city)	Deforestation rate (% of CA per city per year)	Precipitation (mm in CA per month per city)
Rondonia	52	25617	96	0.24	0.008	155
Acre	22	26646	113	0.48	0.007	147
Amazonas	62	44780	157	0.60	0.001	193
Roraima	15	22010	72	0.36	0.005	177
Para	140	45625	32	0.23	0.006	231
Amapa	16	35783	87	0.34	0.002	237
MatoGrosso	94	32043	6	0.24	0.007	156

### 2.2.2 Catchment area

We adopt a 20-km radius as the “catchment area” around each city and use the precipitation, deforestation rate, and forest cover estimates within this catchment area as our covariates. The size of this catchment area accounts for the malaria vector flight range (38, 66), population mobility to and from the surrounding vegetation, and the fact that malaria cases often arise from multiple urban health facilities within a particular city. The same radius has been used elsewhere as the area typically under urban influence in the Brazilian Amazon region (82). Our results are robust to the use of different radii (i.e., 10, 20, and 30 km) (see Figure 23 and additional details in Appendix II).

### 2.2.3 Covariates

Population size comes from the 2007 Brazilian National Census, aggregated at the census tract level, made available by the Brazilian Institute of Geography and Statistics (83). Our environmental covariates come from satellite imagery. We used annual forest cover and deforestation rate estimates from the Brazilian Space Agency derived from a semi-automated analysis of Landsat imagery (84). Estimates of precipitation were derived from the Tropical Rainfall Measuring Mission data ('3B43 Monthly 0.25 x 0.25 degree merged TRMM and other sources estimates' product (85)), and average precipitation for a particular month was calculated over all pixels that fell within each catchment area. Based on these precipitation estimates, we also calculated a drought index that has been extensively used to characterize drought in the region (86-88). We used a one month time lag for precipitation and drought index covariates based on the assumption that water affects the vector mainly through its breeding habitat. Therefore, changes in precipitation or drought should only affect infection risk in the following month since this is the minimum necessary time for the larva to become an adult mosquito, the adult to be infected and finally become infectious. Results did not change substantially when using a two month time lag (data not shown).

### 2.2.4 Permutation tests

To compare a particular outcome  $X$  (e.g., number of malaria cases per month, onwards simply malaria incidence) for cities with characteristic  $c_1$  versus cities with

characteristic  $c_2$  (e.g., high vs. low forest cover), we first calculate the observed difference in means  $Dif^{(obs)} = \bar{X}_{c_1}^{(obs)} - \bar{X}_{c_2}^{(obs)}$ . Then, we estimate the probability of an outcome equal or more extreme than the observed outcome under the null hypothesis (i.e., p-value) through a permutation test. To do this, we randomly assign these characteristics to the cities and calculate the simulated difference in means  $Dif^{(sim)} = \bar{X}_{c_1}^{(sim)} - \bar{X}_{c_2}^{(sim)}$ . This was done 1,000 times, generating 1,000 values of  $Dif^{(sim)}$ . We estimate the p-value as

$$p(Dif^{(sim)} \geq Dif^{(obs)}) \approx \frac{\sum I(Dif^{(sim)} \geq Dif^{(obs)})}{1000},$$

where  $I()$  is the indicator function that

takes on the value of 1 if the condition is satisfied and zero otherwise.

## 2.2.5 Regression model

We assessed the effect of forest cover ( $F_{iy}$ , percent of catchment area) and annual deforestation rate ( $DE_{iy} = F_{iy} - F_{i,y-1}$ , percent of catchment area per year) using a Bayesian hierarchical regression approach ( $i$  and  $y$  stand for city and year). We adjusted for potential confounder effect of climate variables on malaria risk, namely monthly precipitation  $P_{iym}$  and a drought index  $D_{iym}$  ( $m=1, \dots, 12$  stands for month within year). All covariates were standardized (i.e., centered and divided by their standard deviation).

The number of malaria cases per month (i.e., malaria incidence) is modeled as an over-dispersed Poisson, given by:

$$C_{iym} \sim \text{Poisson}(\exp(w_{iym})N_i)$$

where  $N_i$  is the population size within the catchment area, and  $w_{iym}$  is given by:

$$w_{iym} \sim N(\beta_{0i} + \beta_1 DE_{iy} + \beta_2 F_{iy} + \beta_3 P_{iym} + \beta_4 D_{iym} + e_{im} + e_{iy}, \sigma^2)$$

where  $\beta_1, \dots, \beta_4$  are fixed-effect regression parameters. Additional socio-economic-environmental covariates (e.g., proportion of migrants, age distribution, level of urbanization, gross domestic product, vector ecology, and proximity to large water bodies) tend to be relatively constant within the short time-frame of our malaria incidence dataset (~4.5 years). Therefore, we control for these unspecified city-to-city differences using a city specific random intercept ( $\beta_{0i}$ ). We also include a year-by-city ( $e_{iy}$ ) and a month-by-city ( $e_{im}$ ) random effect. To complete the model specification, we adopt the usual assumptions regarding the distribution of random effects:

$$\beta_{0i} \sim N(\beta_0, \tau^2)$$

$$e_{iy} \sim N(0, \varphi^2)$$

$$e_{im} \sim N(0, \gamma^2)$$

and we assume vague hyper-priors for the regression coefficients and variance parameters (89):

$$\beta_0, \dots, \beta_4 \sim N(0, 100)$$

$$\varphi, \sigma, \tau, \gamma \sim Unif(0, 100)$$

We used a Gibbs sampler to iteratively sample from each of the full conditional distributions. Because of conjugacy between the normal distribution and priors, almost

all the parameters could be sampled directly (90). The only parameters that could not be sampled directly were the  $w_{iym}$ , which were sampled with a Metropolis-within-Gibbs step. We assessed model convergence by running three Markov Chain Monte Carlo (MCMC) chains with over-dispersed initial parameter values for 200,000 iterations. We discarded the first 10,000 iterations as burn-in and retained 500 iterations, systematically sampled from the remaining 190,000 iterations. We visually assessed convergence by overlaying trace plots of these three chains. We also assessed convergence by calculating the convergence statistic R suggested by Gelman and Rubin (91) (Table 14 in Appendix II), where R values much greater than 1 indicate lack of convergence. Both, our plots and the convergence statistic R, suggest that convergence has been achieved.

To determine whether our model was over-fitting the data, we performed a validation exercise. In this exercise, we compared the out-of-sample predictive ability of our model versus simpler versions of it, either without the month-by-city random effects  $e_{im}$  or without the year-by-city random effects  $e_{iy}$ . We fitted these models to 90% of the data and used the estimated parameters to predict the 10% of the data that was left out. The results from the validation exercise (data not shown) and the comparison between the data and the predictive posterior distribution for each city (Figure 24) revealed that our model had an adequate fit. Finally, a preliminary analysis indicated that the assumption of a linear relationship between LULC covariates and malaria incidence was adequate and revealed low temporal and spatial correlation (correlation on Pearson

residuals  $< 0.2$ ), suggesting that additional nonlinear terms and parameters to model these correlations were not warranted. All analyses and figures were created using R (29).

### **2.2.6 Land use / land cover (LULC) future scenarios**

To evaluate the long-term effect of conservation strategies in the Amazon basin, Soares-Filho et al. (92) simulated a governance (GOV) scenario and compared it to a business-as-usual (BAU) scenario, revealing that a substantial amount of deforestation (and its deleterious effects) could be avoided. These projections also allow us to evaluate the effect of future LULC trends on malaria. We estimated the ratio of the expected malaria incidence under the GOV scenario  $E(C_{iy}^{GOV})$  and under the BAU scenario  $E(C_{iy}^{BAU})$  for each year and city. This ratio  $E(C_{iy}^{GOV}) / E(C_{iy}^{BAU})$  was calculated using the posterior distribution for the over-dispersed Poisson regression parameters, thus fully accounting for their uncertainty (93).

## **2.3 Results**

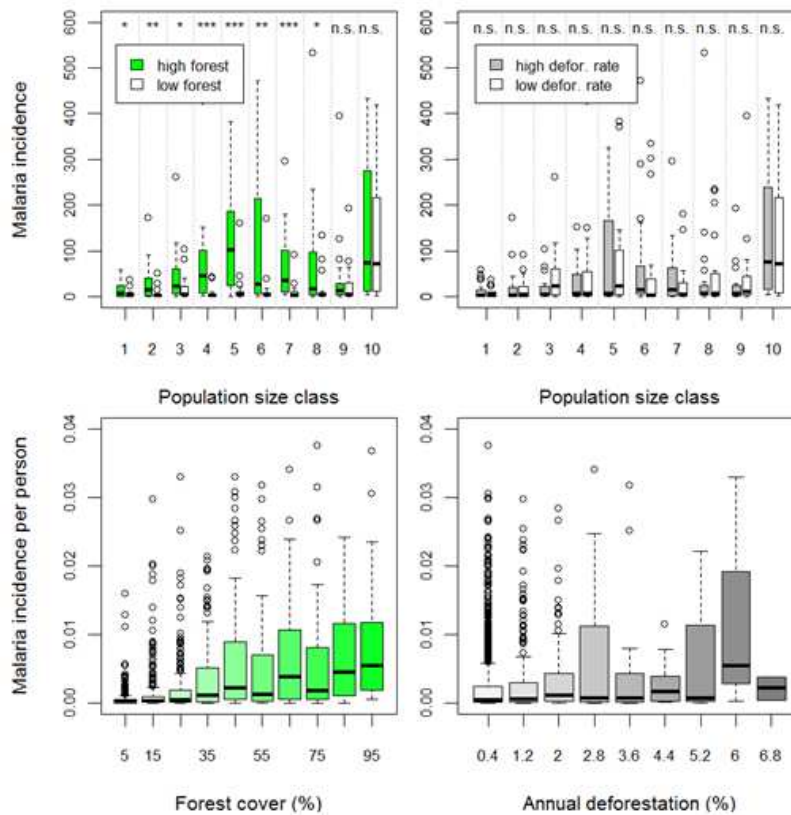
We find overwhelming evidence that areas with higher forest cover tend to be associated with higher malaria incidence whereas no clear pattern could be found for deforestation rates, when comparing cities with similar population sizes (upper panels in Figure 10). Similar evidence arises when analyzing malaria incidence per person across all cities (lower panels in Figure 10). Using a Hierarchical Bayesian regression, we show that although forest cover and deforestation rate were both positively associated

with malaria incidence, forest cover effect was ~25 times greater than that of deforestation rate (Table 5). As a result, the net effect of higher deforestation rates is to decrease malaria burden by decreasing forest cover (i.e., increasing the distance to forest fringes). We also find that the number of malaria cases was negatively correlated with precipitation and our drought index, suggesting that drier periods of the year tend to result in higher malaria incidence. These results were robust to alternative definitions of catchment area (see Appendix II). An alternative model specification, which explores changes in malaria incidence within each city (rather than within and between cities), reveals qualitatively similar results in relation to the LULC variables (Table 13 in Appendix II).

**Table 5: Summary of regression parameter estimates**

LCI and UCI: lower and upper limit of the 95% credible interval

Parameter	Covariate description	Mean	LCI	UCI
$\beta_1$	Annual deforestation rate	0.04	0.00	0.08
$\beta_2$	Forest Cover	1.03	0.87	1.20
$\beta_3$	Precipitation	-0.07	-0.08	-0.05
$\beta_4$	Drought index	-0.06	-0.07	-0.04



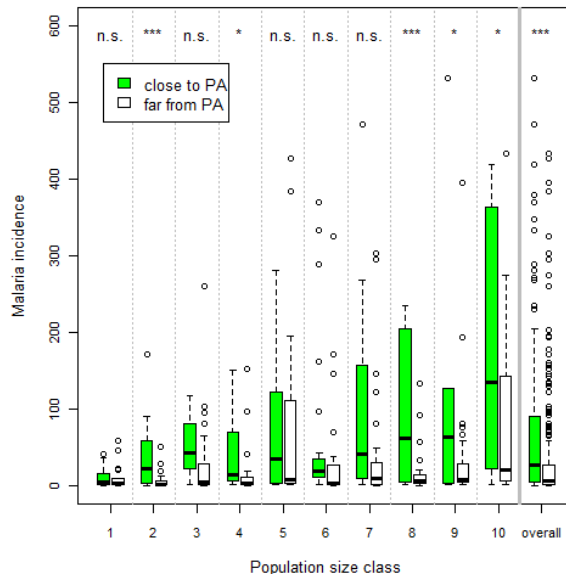
**Figure 10: Malaria incidence is higher in areas with more forest cover whereas no clear pattern arises regarding deforestation rates.**

Upper panels: Data were stratified into 10 percentile population size classes and average number of malaria cases per month for each year and city was depicted. Within each size class, we compare cities with high (green box-plots) vs. low forest cover (white box-plots) (upper left panel); and cities with high (grey box-plots) vs. low deforestation rate (white box-plots) (upper right panel). Cities with high forest cover (or high deforestation rates) are cities that have forest cover (or deforestation rate) higher than the median for that size class. 'n.s.', '\*', '\*\*', and '\*\*\*\*' are non-significant ( $p > 0.05$ ), significant ( $0.01 < p < 0.05$ ), very significant ( $0.001 < p < 0.01$ ) and highly significant ( $p < 0.001$ ) difference in means, respectively, based on permutation tests.

Lower panels: Mean number of malaria cases per month for each year and city divided by total population as a function of forest cover (lower left panel) and deforestation rate (lower right panel). Note: y-axes were truncated to enable a clearer depiction of the bulk of the data (i.e., less than 0.5% observations were excluded from these plots).

These findings have important implications regarding LULC policies in the region. For instance, protected areas (PA's) are a cornerstone of current conservation efforts, yet we are unaware of studies that discuss negative health impacts of these PA's

on the local population. A simple depiction of our malaria data suggest that cities close to protected areas (PA's) tend to have higher malaria incidence than cities far from these PA's (Figure 11) after controlling for population size, a direct consequence of higher forest cover in these areas.

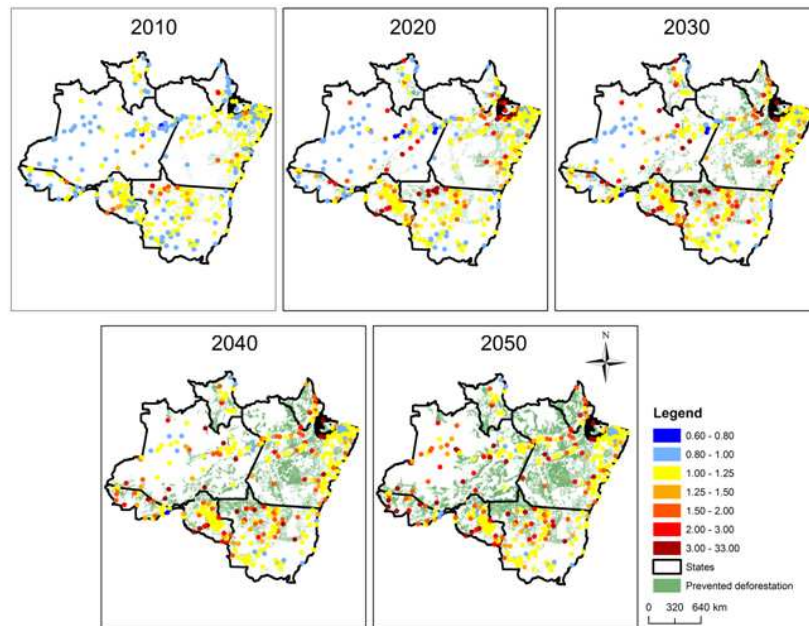


**Figure 11: Malaria incidence tends to be higher for cities close to protected areas (PA's).**

Data were stratified into 10 percentile population size classes and average number of malaria cases per month for each city was depicted. Within each size class, we compare cities close (green box-plots) vs. distant from PA's (white box-plots). Cities close to PA's (i.e., indigenous lands, state and federal parks) are those whose catchment area intersected one or more PA's. 'n.s.', '\*', '\*\*', and '\*\*\*' are non-significant ( $p > 0.05$ ), significant ( $0.01 < p < 0.05$ ), very significant ( $0.001 < p < 0.01$ ) and highly significant ( $p < 0.001$ ) difference in means, respectively, based on permutation tests.

We also evaluated the long-term implications of our findings by comparing a future scenario with reduced deforestation (i.e., governance scenario - GOV) to a future business-as-usual (BAU) scenario. Using our regression parameter estimates, we find that cities with higher malaria incidence in the GOV versus the BAU scenario will

initially tend to be concentrated in the south and east portion of the Brazilian Amazon (Figure 12), where roads slated for paving tend to be located. However, by 2050, almost all cities will tend to have higher malaria incidence.

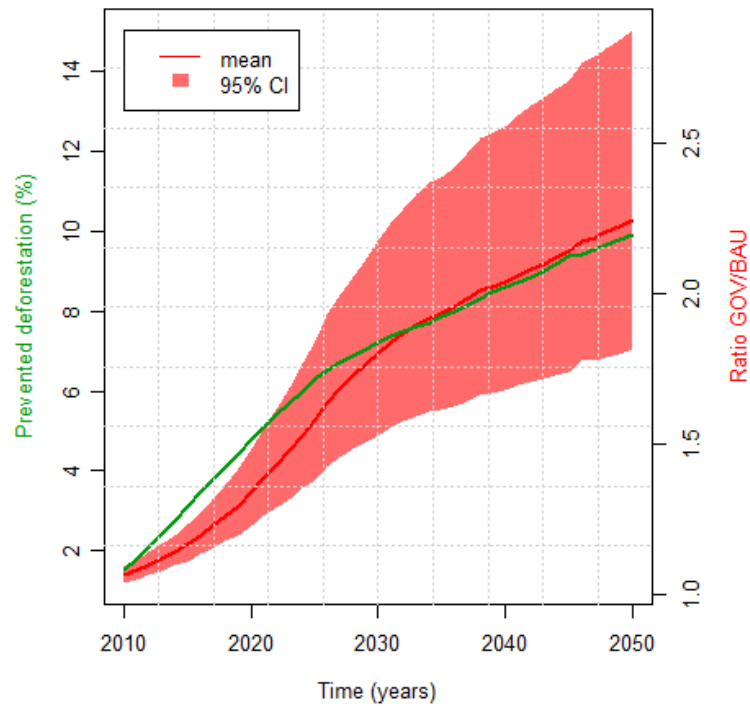


**Figure 12: Predicted malaria incidence in urban health posts is higher in the governance scenario than in the business-as-usual scenario.**

Maps depict the ratio of the expected number of malaria cases per month for each year and city under the governance (GOV) and the business-as-usual (BAU) future LULC scenarios (i.e.,  $E(C_{iy}^{GOV}) / E(C_{iy}^{BAU})$ ), where values  $> 1$  indicate that the GOV scenario results in more malaria cases than the BAU scenario. Areas that were deforested in the BAU scenario but not in the GOV scenario (i.e., prevented deforestation) are depicted in the background for reference. Circles represent the cities in our original malaria dataset.

A summary of these results indicate that avoiding deforestation through better governance can substantially increase malaria incidence in urban health posts; an average of 10% of prevented deforestation resulted in an average 2-fold increase in the

number of malaria cases per month by 2050 (Figure 13). These results raise concern regarding collateral public health effects of conservation policies.



**Figure 13: Malaria incidence increase at urban health posts in the governance scenario is predicted to be a direct consequence of prevented deforestation.**

We depict the relationship between future prevented deforestation under the governance scenario (green line), and the ratio of the expected malaria incidence for each year and city under the governance (GOV) and business-as-usual (BAU) future LULC scenarios (red line) (i.e.,  $E(C_{iy}^{GOV}) / E(C_{iy}^{BAU})$ ), averaged across all cities. The red polygon depicts the 95% credible interval of the mean ratio  $E(C_{iy}^{GOV}) / E(C_{iy}^{BAU})$ .

## 2.4 Discussion

We find that drier periods of the year tended to correlate with higher malaria incidence. Similar results have been attributed to decreased survival rate of adult mosquitoes (94) as well as larva being washed away in rivers (95) during the wet season.

We refrain from further discussing these seasonal patterns here (we will address them in a separate paper) to focus our discussion on the LULC findings.

Malaria risk at frontier regions in the Amazon region is often observed to follow a peculiar time trajectory; in the early phases of human settlement, the number of malaria cases soars as naïve settlers arrive and engage in forest related extractive activities, living in precarious conditions. At later stages, as deforestation increases the distance of settlers to forest fringes and economic conditions improve, malaria risk tends to decrease through time (8). Our findings regarding the LULC covariates agree with this later stage, suggesting that conservation efforts to decrease deforestation in places where people are already settled might inadvertently increase the number of malaria cases. Some would argue that conservation efforts will also decrease the amount of forest related extractive activities (e.g., fishing, hunting, extraction of non-timber forest products), thus decreasing malaria risk. We are skeptical; even if forest conservation efforts succeed in retaining forest cover, hunting and fishing is likely to continue to occur, even within protected areas (e.g., 96, 97, 98).

We note that our finding directly contradicts the growing body of literature that suggests that forest conservation can decrease disease burden (71-76). This literature often cite the study of Vittor et al. (38, 40) conducted in the Peruvian Amazon, as an example of how deforested areas favor the main malaria vector, *Anopheles darlingi*. However, similar entomological studies in the Brazilian Amazon region suggest the

opposite pattern for the same vector species (31, 67, 99, 100), strongly supporting our results. This conflicting evidence might be due to distinct LULC patterns in these regions. In the Peruvian Amazon, swidden-fallow agriculture is the primary driver of deforestation and, as a result, deforested areas are often covered by shrubs and secondary vegetation growth (38, 40), whereas in the Brazilian Amazon, forests tend to be substituted by pasture and soy plantations (101).

Malaria occurring in urban areas is often attributed to poor housing and drainage conditions of slums (e.g., 102, 103). Furthermore, because slums are often located at the periphery of cities and thus closer to forests, this may give rise to a spurious association between forests and malaria incidence. We believe this hypothesis does not explain the malaria patterns we find in the Brazilian Amazon for several reasons. First, slums are rare in Brazilian Amazon cities because these cities are typically very small (i.e., as mentioned earlier, the median population size is 14,000) whereas slums tend to occur in bigger cities where a growing population in limited space gives origin to dense housing, often in hazardous sites. Using the Brazilian government census from 2010, we find that only 12% of the cities in our analysis had slums and that our results in Figure 10 do not change substantially after we exclude the cities with slums (data not shown). Second, the slum effect hypothesis predicts higher malaria incidence in bigger and poorer cities, contrary to the results depicted in Figure 10 and Figure 25 in Appendix II. Finally, even after taking into account gender imbalances in the population

of each city, we find that the average number of malaria cases per month per person tends to be higher in men than in women, a phenomenon that occurred in 96% of the cities in our dataset. This gender difference in malaria incidence agrees with our hypothesis that forest related activities in the surrounding areas, mostly conducted by men, are the cause of higher malaria rates rather than housing conditions.

Unfortunately, policies that have large effect on LULC in the region (e.g., road opening/paving, creation of rural settlement areas, and the establishment of protected areas) are traditionally perceived to lie in the realm of the Ministries of Environment, Infra-structure, Agriculture, and/or Energy, while the Ministry of Health typically focuses on the delivery of health services (7). Similarly, global efforts are typically compartmentalized into conservation (e.g., REDD+) and public health (e.g., Roll Back Malaria and GFATM) initiatives. Few studies identify, or discuss how to address, trade-offs between these global efforts and governmental policies, probably because of the interdisciplinary nature of these trade-offs and the associated ethical issues. For instance, how can one reconcile potential conflicts between the Millenium Development Goals (e.g., goal of combating malaria and the goal of ensuring environmental sustainability)? Although we do not have an answer to this question, acknowledging that these tradeoffs exist is a critical first step towards finding a solution.

Current research and resulting policy recommendations regarding LULC in the Amazon ignore potential public health impacts. The most frequent policy action to

decrease deforestation rates is to create protected areas (104-106). Several studies suggest, however, that many of these protected areas are established in areas with small deforestation risk (105, 107), effectively averting few of the impacts of deforestation. These observations have resulted in recommendations to place these parks in areas more prone to deforestation (59, 107, 108), which often imply areas with larger human populations, disregarding the potential for increased malaria morbidity for the local population. Similarly, research acknowledging the negative effects of conservation efforts typically emphasize restrictions on agricultural development rather than the detrimental impact on public health (106, 109, 110).

One possible interpretation of our findings is that we are promoting deforestation. This is not the case. For instance, large-scale settlement projects in heavily forested areas have resulted in substantial deforestation *and* major malaria outbreaks in the past (8). Here we argue that deforestation has both negative and positive effects in places where people are already settled, and that the knowledge of these effects is essential for proper LULC and public health planning, particularly in light of the recent ambitious REDD+ targets set by the Brazilian government and four of the Brazilian Amazon states (59). If conservation efforts (e.g., REDD+) are to avoid this rapid land cover change and its associated adverse effects on several regional-global environmental services (e.g., atmospheric carbon emission, climate and biodiversity), these conservation efforts should, at a minimum, include proper malaria mitigation strategies

(e.g., creation of more malaria detection and treatment outposts, distribution of long-lasting insecticidal bed nets, indoor residual spraying) to alleviate their local detrimental effects. Similarly, opportunity costs of reduced carbon emissions through conservation initiatives should take into account their local impact on malaria burden.

Our study has five important limitations. First, we do not take into account potential differences between cities in terms of main malaria vector species, vector ecology and infection efficiency. However, it is well known that collection of entomological data is extremely laborious (111) and therefore logistically impossible to collect over the same geographical scale as our malaria data. Yet, finding the same overall result over such a vast area by using a separate regression for each city (Appendix II, model results in Table 13) gives us confidence that our results are robust to these potential city-to-city differences. Second, in the absence of spatial coordinates of the individual health facilities, we rely on data from urban health facilities aggregated at the city level. Yet, we note that even if individual level data had been available, we would still not have been able to consider many individual-level factors that are known to be important for malaria risk (e.g., mobility, socio-economic status, housing conditions, and occupation) because only a few basic demographic characteristics, such as age and gender, are routinely collected by the malaria surveillance system.

Third, in the absence of detailed information for a more accurate modeling of catchment area (e.g., network of unofficial roads (112), origin and mode of

transportation of patients, treatment seeking behavior), we relied on relatively arbitrary radii to delimit the catchment area. Fortunately, our results were robust to changes in these radii. We emphasize that these three limitations are typical limitations of studies conducted over large geographical scales (e.g., the area of a single Brazilian Amazon state, Para, is equivalent to the combined area of France and Spain), illustrating the inherent tradeoff between local detail-rich studies, whose results may or may not be generalizable to a wider region, and large-scale detail-poor studies, which reveal broad scale relationships while ignoring many of the local complexities in malaria transmission. Importantly, while site-specific studies have been critical in shaping our knowledge regarding malaria in the region, they may be ill-suited to evaluate the effect of land clearing versus forest cover because these covariates are often spatially correlated at this scale (i.e., land clearing often occurs in areas with high forest cover). On the other hand, over a large spatial scale, land clearing and forest cover are not highly correlated, allowing us to separately evaluate their effects.

The fourth limitation of our study is that, to avoid spatial extrapolation, our future scenario analysis only considers what would happen to malaria incidence in areas close to where humans are already settled (i.e., the vicinity of urban areas). In these areas, we assume that forests will give place to low intensity cattle ranching and soybean plantations (113, 114), thus increasing the distance between people and forest fringes. On the other hand, had we considered new human settlements (e.g., due to

human migration to new agricultural frontiers), the BAU scenario might have indicated an initial higher malaria incidence due to an initial decrease in distance to forest fringes. Finally, as with any simulation study, our simulation results critically depend on the implicit assumption that everything else (e.g., age distribution, migratory patterns, patterns of natural resource extraction, climate, etc.) remains constant.

The clear pattern in the data (Figure 10 and Figure 11), the consistency of our findings using alternative model specifications, and the evidence from detailed entomological and epidemiological studies in the region (8, 31, 65-67, 99, 100), suggest that the association between forest cover and malaria incidence we found is not spurious. Indeed, vegetation management has long been an important strategy to reduce the incidence of malaria (77). Here we a) show that the effect of forest cover substantially outweighs the effect of deforestation rate (the often cited culprit for malaria in the region) and other climatic variables with a malaria dataset spanning an unprecedented geographical scale; and b) discuss the large-scale multi-sector (i.e., public health, development, and conservation) implications of these findings. Our results suggest caution regarding the widespread assumption that pristine ecosystems will always have beneficial effects for human health (e.g., 71, 72-76, 115-117). We believe there are undoubtedly numerous ecosystem services from pristine environments; however, ecosystem disservices also exist and need to be acknowledged. Coordinated actions from apparently disparate science fields (e.g., epidemiologists and environmental

scientists), government ministries (e.g., Ministry of Health and Ministry of Environment), and the ongoing multi-million dollar conservation and public health efforts in the region, will be required to decrease malaria toll in the region while preserving these important ecosystems.

## **2.5 Citation**

This article has been accepted for publication and should be cited as “Valle, D.; Clark, J. 2013. Conservation Efforts May Increase Malaria Burden in the Brazilian Amazon. PLoS ONE 8(3): e57519”

### Chapter 3. Abundance of water bodies is critical to guide larva control interventions and predict disease risk

Studies on mosquito breeding sites typically survey water bodies to determine larva presence or abundance. Then, measures of association are estimated (e.g., regression coefficients, correlation coefficients, analysis of variance, or t-tests) and used to identify important predictors of larva presence or abundance, with the goal of guiding larva control interventions and predicting disease risk. Entomological studies that follow this generic recipe have been repeatedly conducted across the world for a number of mosquito-borne diseases (a small sample of these studies is given in Table 6). While these measures of association are important to characterize larval habitat, here we contend that these measures may not be enough to guide larva control initiatives and determine disease risk.

**Table 6: Sample of studies on mosquito breeding habitat.**

Vector	Disease	Country	Source
<i>Anopheles gambiae</i>	malaria	Kenya	(118)
<i>Anopheles</i> sp.	malaria	Kenya	(119)
<i>Culex</i>	filariasis and arboviruses	Kenya	(120)
<i>Anopheles</i> sp.	malaria	Kenya	(121)
<i>Anopheles</i> sp.	malaria	Ethiopia	(122)
<i>A. gambiae</i>	malaria	Ghana	(123)
<i>Anopheles</i> sp.	malaria	Tanzania	(124)
<i>Anopheles</i> sp.	malaria	Côte d'Ivoire	(125)
<i>Anopheles darlingi</i>	malaria	Peru	(38)
<i>Anopheles sinensis</i>	malaria	China	(126)
<i>Aedes</i> sp. / <i>Anopheles</i> sp.	dengue / malaria	Thailand	(127)

<i>Aedes aegypti</i>	yellow fever / dengue	Argentina	(128)
----------------------	-----------------------	-----------	-------

Our contention is based on the same arguments as those that motivated the creation of the population attributable fraction (PAF) concept. In the case of PAF, it has been argued that measures of association do not take into account the prevalence of the different risk factors. Thus, a particular risk factor might be statistically significant but have small public health relevance if very few people have that risk factor (129).

Similarly, everything else being equal, the risk factors associated with a productive larval habitat (defined here as a water body that typically has larvae or in which larvae are abundant) might not be relevant if water bodies with those risk factors are rare in the overall landscape. As an over-simplified example, say we have 100 water bodies within the same distance from a village. One water body has characteristics that lead us to predict it surely has larva while 50 water bodies are predicted to have a 25% chance of having larva. It is often easier and also more effective to identify and treat these 50 water bodies than to keep searching for the one water body which certainly has larva.

Determining the relative abundance of water bodies is also critical when predicting disease risk. The key to understand this statement is the fact that researchers typically perform their analysis given that a water body was sampled. Thus, if we want to understand disease risk, we also have to know how the abundance of water bodies changes with the different covariates. In statistical terms, the typical analysis makes inference on the conditional probability  $p(L|W)$ , where L denotes the presence (or

abundance) of larva and  $W$  denotes the event that a water body was sampled. On the other hand, to understand disease risk we need to make inference on the marginal probability  $p(L)$ .

To illustrate, consider another simplified example, summarized in Table 7. We are interested in the risk of malaria infection for a person living in a forested site versus living in a deforested site. Thus, we sample water bodies in both sites using the same number of transects, all of equal length. In scenario 1, these transects yield 30 water bodies (8 of which had larva) in the forested sites and 10 water bodies (8 of which had larva) in the deforested site. As a result, the proportion of water bodies with larva in the forested site is  $\hat{p}_{for} = 8/30 \approx 0.27$  while for the deforested site it is  $\hat{p}_{def} = 8/10 = 0.8$ . Based on these probabilities, a logistic regression would indicate that forest cover is negatively associated with presence of the malaria vector larva. Armed with these results, a researcher might conclude that people living at forested sites have a lower infection risk. This conclusion is incorrect, since both sites have the same number (i.e., 8) of water bodies with larva per area. If these sites give rise to a similar number of larvae and adult mosquitoes, and if these mosquitoes have the same degree of contact with the host, then infection risk should be similar. Alternatively, scenario 2 assumes that these transects yield 30 water bodies (15 of which had larva) in the forested sites and 10 water bodies (5 of which had larva) in the deforested sites. In this case, the proportion of water bodies with larva would be identical in both sites  $\hat{p}_{for} = \hat{p}_{def} = 1/2$  and a logistic

regression would fail to find significant differences between sites. With these results, the researcher would conclude that there is no difference in malaria risk for people living in the forested versus deforested sites. Again, this conclusion is incorrect because the forested site has three times more water bodies with larva per area when compared to the deforested site, probably resulting in a higher number of adult mosquitoes and higher malaria risk in the forested site. In Appendix III, we provide simulation results involving logistic and Poisson regressions, for larva absence/presence data and for larva abundance data, respectively.

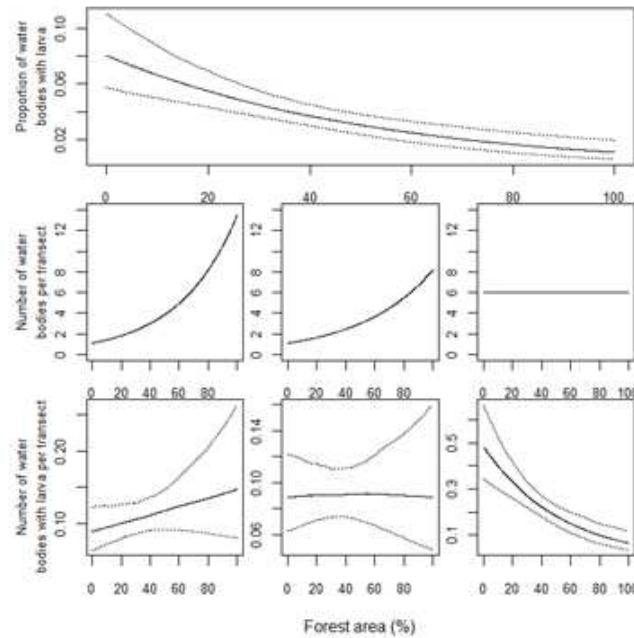
**Table 7: Description of outcomes for scenarios 1 and 2.**

Outcomes	Scenario 1		Scenario 2	
	Forested	Deforested	Forested	Deforested
# of water bodies	30	10	30	10
# of water bodies with larva	8	8	15	5
Proportion of water bodies with larva	0.3	0.8	0.5	0.5

This example is clearly over-simplistic for at least three reasons. First, the size and biochemical composition of water bodies, for instance, may also be considerably different among these sites, leading to higher mosquito productivity in one site versus the other. Second, it might be more difficult to detect water bodies in forested sites than in deforested sites. Finally, an important assumption is that a higher number of water bodies with larva per area results in a higher density of adult mosquitoes and thus a higher disease risk. While this is often a reasonable assumption, there are a number of other factors that also need to be taken into account when determining disease risk (e.g.,

man-biting rate of adult mosquitoes, the proportion of infected mosquitoes [sporozoite rates], and distance to where people live). Nevertheless, the example is useful to highlight that, given everything else being equal, it is important to identify the characteristics of productive larva sites and to take into account the prevalence of water bodies with these characteristics.

To illustrate this problem with real data, we estimated how the proportion of water bodies with larva changes with forest cover for *Anopheles darlingi*, the primary malaria vector in the Amazon region, using data from the Peruvian Amazon. Details regarding how data were collected are available elsewhere (38). We fitted a binomial model (model details are given in Appendix III), which revealed that the probability of a water body having *A. darlingi* larvae decreases with forest cover (upper panel in Figure 14). However, depending on the relationship between the number of water bodies per transect and forest cover (middle panels in Figure 14), the relationship between the number of water bodies with larva per transect and forest cover can vary substantially and even reverse sign (lower panels in Figure 14).



**Figure 14: The number of water bodies with larva per transect (lower panels) is influenced by the relationship between the proportion of water bodies with larva and forest area (upper panel) *and* the relationship between number of water bodies per transect and forest area (middle panels).**

Solid and dashed lines are median and 95% credible intervals, respectively.

We believe it is critical for researchers to carefully consider how the outcome of their analysis will inform policy actions. The typical regression analysis assumes water bodies to be the sampling unit, thus yielding results per water body. On the other hand, if the researcher is primarily interested in infection risk, it is likely that the response variable more closely associated with infection risk is in areal unit (e.g., number of larva per transect or number of water bodies with larva per transect). In other words, there is a mismatch between the analyzed outcome and the outcome more relevant for public policy. To avoid this mismatch, we propose two alternatives. First, one can directly

model the number of larva per transect (or number of water bodies with larva per transect) as a function of transect-level covariates, assuming that the sampling unit is the transect. Alternatively, one can predict the number of water bodies per transect to then predict the mean larva abundance (or larva presence) of these water bodies (as in Figure 14). Both approaches could be also used for fixed-area plots or houses (e.g., for vectors found in peridomiliar areas, such as *Aedes aegypti*; 130). In either modeling approach, the sampling design for water bodies is critical and merits careful consideration.

Unfortunately, most studies provide detailed description of how larvae were sampled within water bodies but not how water bodies themselves were sampled (e.g., 122). It remains an important research area to determine how water body abundance changes as a function of landscape characteristics.

An important problem in using the individual transects as sampling units is that water bodies may be very heterogeneous within a transect. To some extent, this can be circumvented by having shorter transects but there is an inherent trade-off between transect length and data collection effort (i.e., it is harder to create 10 transects of 100 m than a single transect of 1,000 m). This same trade-off exists for plots as well. We also note that, in some cases, the notion of discrete water bodies does not make sense. For instance, a researcher might be interested in determining larval productivity of different types of wetlands or rivers. Yet, we still believe that it is not enough to determine the productivity of these potential larva habitats to be able to infer malaria risk; one must

know the proportion of the area covered by these different habitats as well. We emphasize that using water bodies as the sampling unit in statistical analysis is perfectly valid to characterize larval habitat. However, researchers should be careful when using the derived measures of association to identify larval control strategies and predict disease risk. While dengue researchers have long recognized the importance of accounting for the abundance of water containers (e.g., 131), we believe that entomological researchers focused on other diseases are largely unaware of the issues we raise. Finally, although we have focused on mosquito larva habitat, our results are likely to apply to other types of disease vectors that also rely on water bodies.

## **Chapter 4. Revisiting the relationship between land cover and malaria in the Peruvian Amazon: the importance of integrating data on multiple vector species, mosquito life stages, and malaria prevalence**

### **4.1 Introduction**

It has been widely acknowledged that the environment strongly influences disease vectors and thus vector-transmitted diseases (3). In particular, the influence of land use / land cover changes on malaria in the Amazon region has attracted considerable attention because this region concentrates most of the malaria cases in the Americas (7, 60-62) as well as large fraction of worldwide tropical deforestation (57). In 2001, a landmark study was conducted in the Peruvian Amazon to examine how larva breeding habitat and mosquito biting rate are influenced by land use (38, 40). Based on *A. darlingi* data, this study concluded that deforestation increases malaria risk and since then it has been widely cited as an example of a win-win situation for conservation and public health (e.g., 72, 73, 75). Other entomological and epidemiological studies, however, have suggested that malaria risk is actually higher in forest fringes but not in highly deforested landscapes (8, 31, 65-67). For instance, a recent analysis of a large malaria incidence data from the Brazilian Amazon revealed that the long-term effect of deforestation is to increase the distance to forest fringes, leading to lower malaria risk (132).

Here we revisit the original dataset from the Peruvian Amazon with the goal of illuminating the causes of this controversy. We build upon the original manuscripts in two aspects. First, we examine multiple malaria vector species simultaneously rather than just *A. darlingi*. This is important because, despite the frequent focus on *A. darlingi* as the primary malaria vector (38, 40, 133), several other anopheline species have been shown to be competent vectors, such as *A. oswaldoi*, *A. nuneztovari*, *A. rangeli*, *A. benarrochi*, and *A. triannulatus* (133-136). As a consequence, inference on malaria risk based solely on data from *A. darlingi* might not necessarily be valid. Second, we integrate mosquito data with malaria prevalence data. Higher malaria vector abundance might suggest, but does not necessarily imply, higher malaria risk in certain areas (e.g., paddies paradox; 137).

## **4.2 Methods**

### **4.2.1 Data**

This study was conducted along the Iquitos-Nauta road in the Peruvian Amazon between 2000 and 2001. A detailed description of the area (e.g., environmental conditions, land use, road construction, deforestation patterns, and settlement history) can be found in Marki et al. (138).

The mosquito data contained information on larva abundance and mosquito biting rate. Larval anophelines were collected once every three weeks between March and September 2001 from water bodies along multiple transects. On total, ~3,000 water

bodies were examined, from which 1,394 contained larvae from one or more anopheline species. Details regarding the larval collection methodology can be found in (38, 139). Adult anophelines were captured between 18:00 and 24:00 using human bait. This was done once every three weeks between August 2000 and August 2001 over 56 sites, resulting in the collection of ~11,500 anopheline mosquitoes. An important aspect regarding data collection is that areas close to rivers were explicitly avoided by reasoning that vector ecology in these places would be substantially different from vector ecology in tierra firme. More information regarding the adult collection methodology can be found in (40, 139). Throughout this article, we abbreviate species names as *A. darlingi* (dar), *A. nuneztovari* (nun), *A. triannulatus* (tri), *A. benarrochi* (ben), *A. oswaldoi* (osw), and *A. rangeli* (ran) in our figures.

Malaria prevalence data come from a cross-sectional survey conducted between July and August 2001. A total of ~2,800 individuals distributed across 18 villages had their blood sampled, regardless of symptoms, and approximately 97% of the microscopy slides were examined more than once to confirm diagnosis. Basic demographic information as well as data on behaviors that might be relevant for malaria infection were also collected. Overall, 115 (3.7%) individuals were detected to have malaria by at least one microscopist. Additional details regarding malaria data collection are available in (139).

### 4.2.2 Regression models

Our approach to analyze larva, mosquito biting rate, and malaria prevalence data was to create multiple Bayesian regression models. We start by creating a model to predict the abundance of water bodies with larva for each species. Using predictions from this model as an additional covariate, the second model focused on the mosquito biting rate from the same species. Finally, using predictions of mosquito biting rate as covariates, we determine the factors that influence malaria prevalence. We created customized models to properly accommodate data idiosyncrasies, such as zero-inflation and over-dispersion. Customization also allows us to generate more parsimonious predictive models, achieved by using multiple shrinkage priors on the slope parameters and a reversible jump algorithm. Each of these models was fitted using a Gibbs sampler and Metropolis-within-Gibbs steps whenever full conditional distribution were not available in closed form. Algorithm convergence was assessed using trace-plots and model fit was evaluated by comparing data to the corresponding predictive distributions. For brevity, we provide a description of the likelihood here but we leave the description of our priors and model fit to Appendix IV. In all these models, posterior slope estimates  $\beta$  for which  $\min(p(\beta < 0), p(\beta > 0)) < 0.025$  were deemed statistically significant.

#### 4.2.2.1 Larva model

Let the number of water bodies with larva of the  $s$ -th species in the  $j$ -th transect at time  $t$  be denoted by  $L_{sjt}$ . For the two most abundant species, we model  $L_{sjt}$  using a zero inflated binomial model. In other words, we assume:

$$L_{sjt} | \omega_s, p_{sjt} \sim \omega_s + (1 - \omega_s) \text{Binomial}(0 | W_{jt}, p_{sjt}) \text{ if } L_{sjt} = 0$$

$$L_{sjt} | \omega_s, p_{sjt} \sim (1 - \omega_s) \text{Binomial}(L_{sjt} | W_{jt}, p_{sjt}) \text{ if } L_{sjt} > 0$$

where  $\omega_s$  is the zero-inflation probability,  $W_{jt}$  is the number of water bodies, and  $p_{sjt}$  is the success probability of the binomial distribution. For the remaining less abundant species, we employ a regular binomial model by assuming that  $\omega_s = 0$  because a preliminary analysis revealed identifiability problems regarding  $\omega_s$ . Finally, we assume that  $\text{logit}(p_{sjt}) = \alpha_{1s} + \mathbf{x}_{jt}^T \boldsymbol{\beta}_{1s}$ , where  $\mathbf{x}_{jt}$  is the design vector containing LULC and climate covariates,  $\alpha_{1s}$  and  $\boldsymbol{\beta}_{1s}$  are the species specific intercept and vector of slopes, respectively.

#### 4.2.2.2 Mosquito biting rate model

Here we adopt a zero-inflated Negative Binomial model, where zeroes are interpreted to have two origins: not enough sources of young adults or low adult mosquito survival. Let the zero-inflation probability for location  $j$  at time  $t$  for species  $s$  be denoted by  $1 - \theta_{sjt}$ . We assume that  $\theta_{sjt}$  (onwards referred to simply as the probability of young adults) depends solely on the predicted mean number of water

bodies with larva from the same species  $\widehat{L}_{sj\tilde{t}}$ , given by  $E[(1 - \omega_s) p_{sj\tilde{t}}] \times \bar{W}$ , where  $\bar{W}$  is the average number of water bodies among all 1,000 m transects. This assumption is formalized as:

$$\text{logit}(\theta_{sjt}) = \alpha_{2s} + \beta_{2s} \widehat{L}_{sj\tilde{t}}$$

where  $\alpha_{2s}$  and  $\beta_{2s}$  are species specific intercept and slope, respectively. The mean of the predicted number of water bodies with larva was assessed one and two weeks prior to the biting rate collection date (i.e.,  $\tilde{t} = t - 7$  or  $\tilde{t} = t - 14$ ) to account for the time needed for the larva to develop and become an adult mosquito.

If a particular location and time has young adults from species  $s$ , we assume that the total number of adult mosquitoes caught  $A_{sjt}$  can be modeled using a Negative Binomial regression, where the mean  $\mu_{sjt}$  is a function of several environmental covariates. The resulting zero-inflated Negative Binomial model can be succinctly described as:

$$A_{sjt} \mid \theta_{sjt}, \mu_{sjt}, r_s \sim (1 - \theta_{sjt}) + \theta_{sjt} NB(0 \mid \mu_{sjt}, r_s) \text{ if } A_{sjt} = 0$$

$$A_{sjt} \mid \theta_{sjt}, \mu_{sjt}, r_s \sim \theta_{sjt} NB(A_{sjt} \mid \mu_{sjt}, r_s) \text{ if } A_{sjt} > 0$$

where  $r_s$  is the over-dispersion parameter for species  $s$ . Given that there are enough young adults, mosquito biting rate has mean  $\mu_{sjt} = \exp(\alpha_{3s} + \mathbf{x}_{jt}^T \boldsymbol{\beta}_{3s})$ , where  $\alpha_{3s}$  and  $\boldsymbol{\beta}_{3s}$  are the species specific intercept and vector of slopes, respectively. Furthermore,  $\mathbf{x}_{jt}^T$  is the design vector containing LULC and climate covariates as well as population density.

We interpret  $\beta_{3s}$  as indicating how adult mosquito survival is influenced by the different environmental covariates.

A preliminary analysis revealed that these mosquito biting rate data contained a few extremely high values for mosquito biting-rate and that these observations strongly influenced our slope estimates  $\beta_{3s}$ . Two options to avoid this undesirable feature would be to collapse these data into presence/absence data or group the data into low, medium, and high biting-rates categories. Both of these options are likely to result in substantial loss of information. We chose a third alternative, which consisted in censoring the observations that were greater than the 99<sup>th</sup> percentile of biting rate for each species  $s$  ( $q_{99,s}$ ). We believe this approach avoids having our slope estimates overly influenced by a few extreme observations while also avoiding the loss of too much information. To account for this censoring, we modify slightly our model:

$$A_{sijt} \mid \theta_{sijt}, \mu_{sijt}, r_s \sim (1 - \theta_{sijt}) + \theta_{sijt} NB(0 \mid \mu_{sijt}, r_s) \text{ if } A_{sijt} = 0$$

$$A_{sijt} \mid \theta_{sijt}, \mu_{sijt}, r_s \sim \theta_{sijt} NB(A_{sijt} \mid \mu_{sijt}, r_s) \text{ if } 0 < A_{sijt} < q_{99,s}$$

$$A_{sijt} \mid \theta_{sijt}, \mu_{sijt}, r_s \sim \theta_{sijt} \sum_{A_{sijt}=q_{99,s}}^{\infty} NB(A_{sijt} \mid \mu_{sijt}, r_s) \text{ if } A_{sijt} \geq q_{99,s}$$

#### 4.2.2.3 Malaria prevalence

We summarize results from the multiple blood smear examinations into a single binary outcome per individual (i.e., 0 if all results were negative or 1 if at least one result

was positive), which we refer to simply as infection status. The rationale for this procedure is that false positive microscopy results are typically rare (13, 17, 22, 34); thus one or more positive detections provide strong evidence that the person is infected. Let this binary outcome for individual  $i$  be denoted by  $y_i$ , which we model using a probit regression:

$$y_i \sim \text{Bernoulli}(\Phi(\alpha_4 + \mathbf{x}_i^T \boldsymbol{\beta}_4)) ,$$

where  $\Phi()$  is the cumulative distribution function of the standard normal distribution,  $\alpha_4$  and  $\boldsymbol{\beta}_4$  are the intercept and vector of slope, respectively, and  $\mathbf{x}_i^T$  is a vector containing the covariates for individual  $i$ . This design vector includes the predicted biting rate of each adult mosquito species  $\hat{A}_{sjt}$ , given by  $E[\theta_{sjt} \mu_{sjt}]$ , together with basic demographic information (i.e., age and gender) and bed-net usage information (i.e., time entering and leaving the mosquito bed-net as well as total amount of time under the bed-net). These covariates entered the model as linear and quadratic terms. Then, using the latent state representation of the probit regression (see Appendix IV), we are able to perform model selection using the reversible jump algorithm described in Denison et al. (140). All parameters used in our regression models are summarized in Table 8.

**Table 8: Summary of the data and model parameters**

Data	Description	Model
$L_{ijt}$	Number of water bodies with larva	Larva model
$W_{jt}$	Number of water bodies per transect	Larva model
$A_{ijt}$	Mosquito biting-rate	Mosquito biting-rate model
$y_i$	Binary outcome of malaria test	Malaria prevalence model
<b>Parameters</b>		
$\omega_z$	Zero inflation probability for larva model	Larva model
$p_{ijt}$	Probability of finding larva in one water body given $\omega_z = 0$	Larva model
$\alpha_{1z}$	Intercept for $\text{logit}(p_{ijt})$	Larva model
$\beta_{1z}$	Slope vector for $\text{logit}(p_{ijt})$	Larva model
$\theta_{ijt}$	One minus the zero inflation probability for biting-rate model	Mosquito biting-rate model
$\mu_{ijt}$	Mean of negative binomial	Mosquito biting-rate model
$r_z$	Over-dispersion parameter for negative binomial	Mosquito biting-rate model
$\alpha_{2z}$	Intercept for $\text{logit}(\theta_{ijt})$	Mosquito biting-rate model
$\beta_{2z}$	Slope for $\text{logit}(\theta_{ijt})$	Mosquito biting-rate model
$\alpha_{3z}$	Intercept in mean of negative binomial	Mosquito biting-rate model
$\beta_{3z}$	Slope vector in mean of negative binomial	Mosquito biting-rate model
$\alpha_4$	Intercept for the probit regression	Malaria prevalence model
$\beta_4$	Slope vector in probity regression	Malaria prevalence model

#### 4.2.2.4 Covariates

The land use / land cover (LULC) covariates were based on an unsupervised classification of a 2001 Landsat image, resulting in seven classes: clouds, cloud shadow, forest, non-forest vegetation, water, impervious area, and deforested area. We calculated the proportion of the area covered by each category using four different radii (i.e., 100, 250, 500, and 1000 meters) from the center of the larva transects and from each adult mosquito collection site. A preliminary analysis revealed that the proportion of deforested area and forested area were highly correlated within each radius. For this reason, we decided to discard the proportion of deforested area. We also calculated the distance from each adult mosquito collection site to the nearest pixel in each LULC category.

Human population density data are based on a census conducted in 2001, where we mapped all households and quantified the number of people in each one of them. Similar to the LULC covariates, human density was calculated within four radii (100, 250, 500 and 1000 m) from adult mosquito collection sites and we also assessed the distance from each site to the closest house. These LULC and human density covariates were divided into five sets (one for each radius and one for the distance based covariates) and we chose the covariate set that had the highest out-of-sample predictive skill (see validation section below).

In relation to the climate covariates, we had daily estimates of several quantities that could be relevant for vector ecology, such as minimum and maximum air and surface temperature, soil moisture, solar radiation, precipitation, and simulated root zone degree of saturation. A short description of how these covariates were estimated is given in Appendix IV. Because the immediate past of these covariates is likely to be more important than these covariates on the day of data collection, we calculated a five-day average (based on five-days prior to data collection) for each climate covariate. Our preliminary analysis revealed a high degree of correlation among some of these climate covariates (e.g., air and surface temperatures, maximum temperature and solar radiation, root zone degree of saturation and soil moisture), so we decided to retain just minimum and maximum temperatures, precipitation and soil moisture. Because pixel size for these climate covariates was large relative to that of the LULC classification (i.e.,

$\geq 1,000$  versus 30 m, respectively), we did not assess how these covariates changed with different radii. All covariates were standardized to have a mean of zero and a standard deviation of one.

#### **4.2.2.5 Validation**

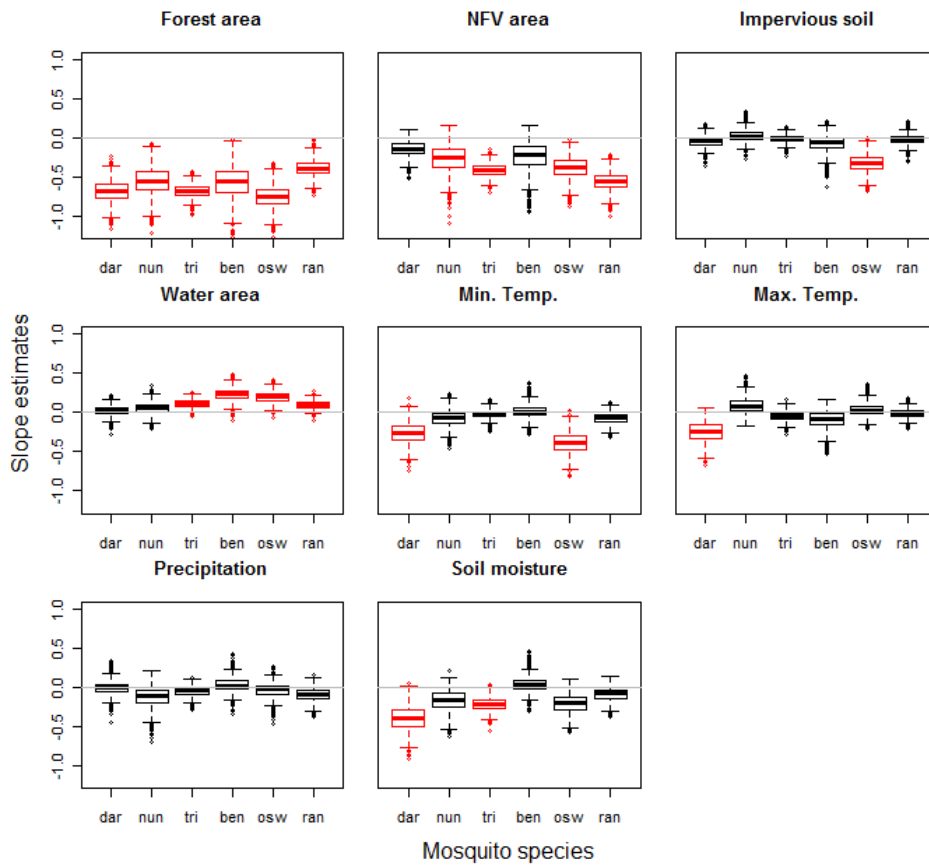
For the biting rate model, we fitted ten distinct models (one model for each combination of the five covariate sets and two larva time lags). On the other hand, because larva transects were up to 1,000 m in length, we reasoned that using distance based covariates or covariates assessed within a radii smaller than 500 m to the center of the transect would not be appropriate. Therefore, for the larva model, we just fitted two distinct models (i.e., with LULC covariates assessed within a radius of 500 m and 1,000 m).

We performed a simple validation exercise to determine the out-of-sample predictive skill of these ten biting-rate models and two larva models. We fitted each model to randomly chosen 90% of the data and evaluated its predictive ability on the remaining 10% of the data. We chose the biting-rate model and the larva model that had the best predictive performance, based on their mean-squared-error (MSE).

## **4.3 Results**

### **4.3.1 Parameter estimates and individual climate and LULC effects**

The larva model with the 500 m radius covariates had a better out-of-sample predictive ability (smaller MSE) when compared to the model with covariates assessed within a 1,000 m radius (data not shown). Thus, from here onwards, we just report results from the 500 m model. The larva model revealed that breeding habitat productivity for all mosquito species tended to be influenced in a similar fashion by climate and LULC variables (Figure 15). Higher forest and NFV area tended to decrease the probability of finding larva for most species whereas proximity to areas with very large water bodies (as detected by remote sensing) tended to increase this probability. The negative effect of minimum and maximum temperatures suggests that larva mortality is increased and/or there is less overall mosquito breeding habitat as temperatures increase. The effect of soil moisture was surprising. Because this covariate takes into account shading, the negative relationship between soil moisture and probability of finding larva might be indicating that shaded water bodies tend to have lower probability of having larva, thus agreeing with the signs of the forest and NFV area slopes. Finally, impervious soil cover was only significant for *A. oswaldoi*.



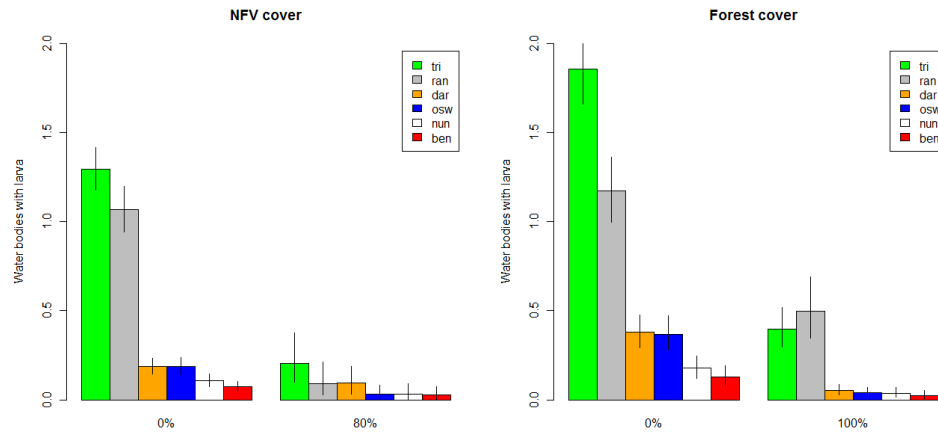
**Figure 15: Posterior distribution of slope parameters, stratified by covariate and mosquito species, for the larva model.**

Red (black) boxes indicate slopes that are (are not) significantly different from zero. A line at zero was added for reference (grey line).

It is particularly noteworthy that the effect size of forest and NFV area was often greater in magnitude than slope estimates for the remaining covariates. Because covariates were standardized, this suggests that forest and NFV area are more important predictors of larva presence than the other covariates. It is also interesting that, among all these different species, *A. darlingi* is the species that is most influenced by climate

(i.e., probability of finding larva from this species was significantly influenced by three of the four climate covariates).

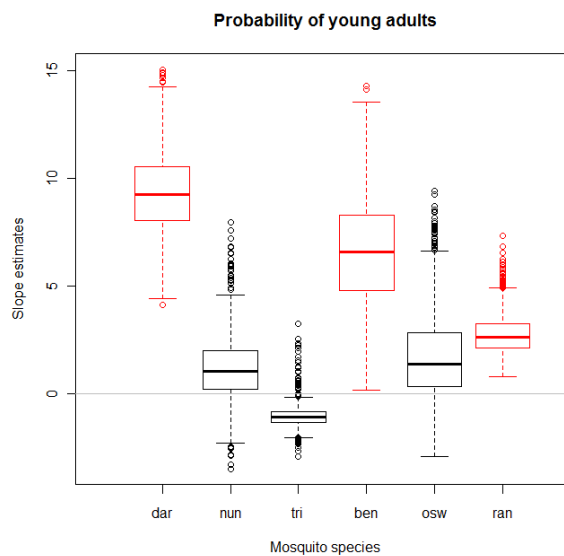
Slope estimates, however, hide important differences in the abundance of water bodies with larva. We find that water bodies with larva from *A. triannulatus* and *A. rangeli* are far more common than water bodies with larva from other species (Figure 16). Furthermore, there is a marked decrease of abundance of water bodies with larva in vegetated (with forest or NFV) areas, with substantial change in species composition. For instance, *A. darlingi* comprises a much larger fraction of the overall larva population in areas with abundant NFV when compared to areas without NFV.



**Figure 16: Abundance of water bodies with larva decreases with forest cover (right panel) and NFV cover (left panel).**

Predicted abundance of water bodies with larva of different species as a function of NFV cover (left panel) and forest cover (right panel). Error bars are 95% credible intervals of the mean predicted abundance. The values for the x-axis were chosen based on the range of the corresponding covariates in the original data.

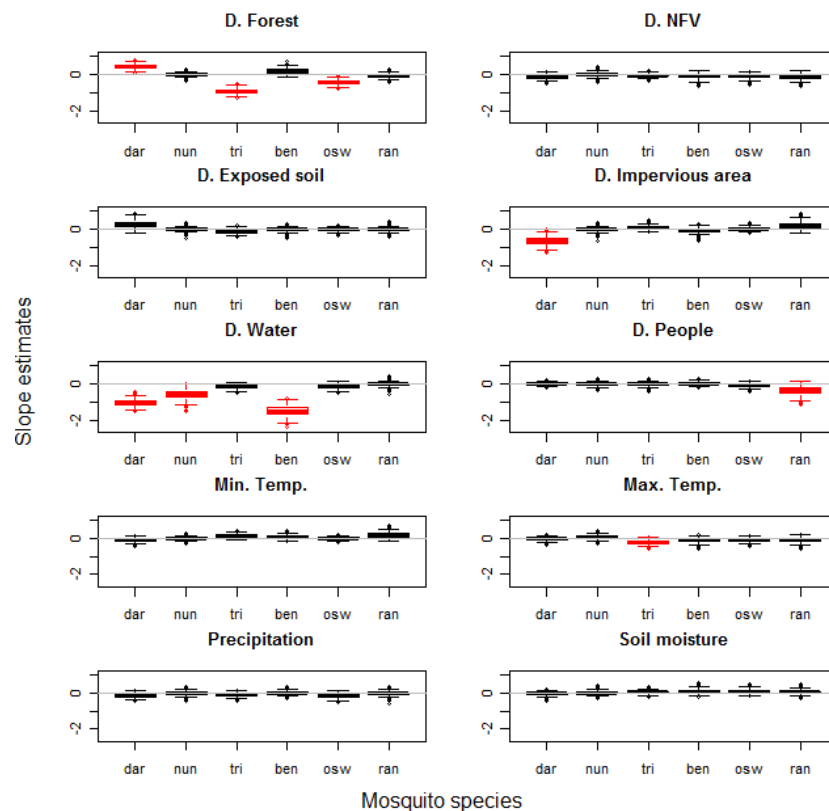
In relation to biting rate, we find that the model with the distance based covariates and two week time lag on the predicted number of water bodies with larva had a better out-of-sample predictive ability (data not shown). As a consequence, we only report results from this model. Our  $\beta_{2,s}$  slope estimates indicate that the predicted number of water bodies with larva in general tended to be positively associated with the probability of finding adult mosquitoes of the same species using human bait, as expected (Figure 17).



**Figure 17: Posterior distribution of slope parameters  $\beta_{2,s}$  showing how mean predicted number of water bodies with larva  $\hat{L}_{sj\bar{t}}$  affects the probability of young adults  $\theta_{sjt}$**

Red (black) boxes indicate slopes that are (are not) significantly different from zero. A line at zero was added for reference (grey line).

Our results also indicate that, given the presence of young adults, mosquito biting rate responds little to climate covariates; the main drivers seem to be distance to forests and large water bodies (Figure 18). The effect of large water bodies are surprising because it reveals an effect that goes beyond breeding habitat, suggesting that these large water bodies may play an important role for adult mosquito survival. Finally, differently from the other mosquito species (in particular *A. triannulatus* and *A. oswaldoi*), *A. darlingi* seem to thrive in more anthropogenic areas, particularly in sites that are farther from forests and closer to impervious area.



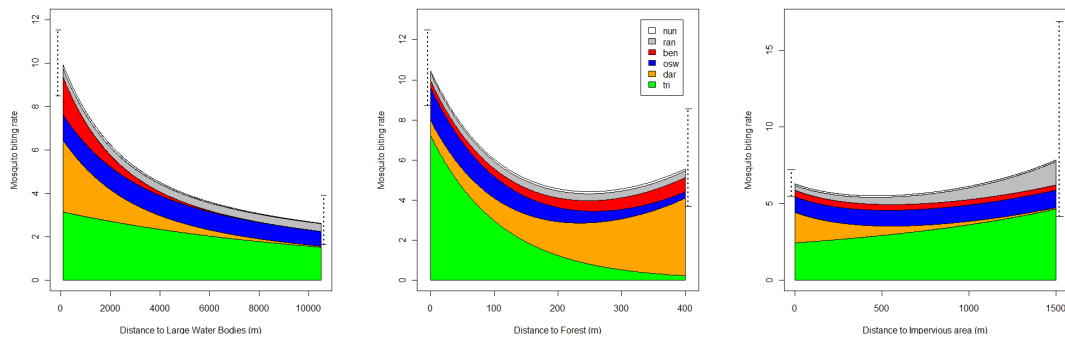
**Figure 18: Posterior distribution of slope parameters  $\beta_{3s}$  showing how biting rate, given larva habitat suitability, is affected by the different covariates.**

Red (black) boxes indicate slopes that are (are not) significantly different from zero, respectively. A line at zero was added for reference (grey line).

Interestingly, the effect of some of our covariates varied considerably for the different mosquito life stages. For instance, forest cover had a negative effect on the number of water bodies with larva for all six mosquito species but the biting-rate of *A. triannulatus* and *A. oswaldoi*, given the presence of young adults, was higher in areas closer to forests. These contrasting patterns may indicate different habitat requirements for the larva versus the adult mosquito life stages. Our results highlight the challenge of inferring how these covariates affect malaria risk; while a particular covariate may increase the probability of larva presence, it may also decrease mosquito biting rate (and vice-versa).

We again find that the most abundant anopheline species was *A. triannulatus*. However, contrarily to the patterns in the larva data, the other two most common adult mosquito species were *A. darlingi* and *A. oswaldoi* while adult mosquitoes from *A. rangeli* were far less abundant than expected (Figure 19). This pattern might be due to the fact that observations were restricted to the period from 18:00 to 24:00. We find that overall mosquito biting rate significantly decreases with distance to the forest and large water bodies. There are also striking changes in species composition; in particular, as distance to forest increases and distance to impervious area decreases, *A. darlingi* becomes increasingly the main adult mosquito species biting humans, clearly revealing how *A. darlingi* is well adapted to human modified environments. On the other hand, we find

that *A. triannulatus* is the dominant vector in areas closer to forests and farther from impervious areas.



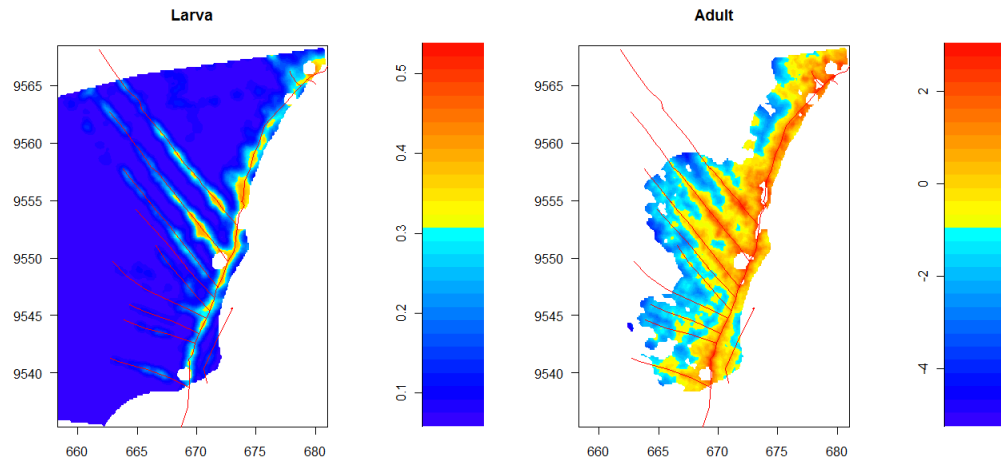
**Figure 19: Predicted mosquito biting rate as a function of distance to large water bodies (left panel), to forest (middle panel), and impervious areas (right panel).**

The limits for the x-axis were chosen based on the range of the corresponding covariates in the original data. Vertical dashed lines depict the 95% credible interval of the predicted total number of mosquito bites over all six species.

### 4.3.2 Spatial patterns

The analysis of the individual LULC effects can be a bit misleading in that these characteristics are not entirely independent throughout the study region. Thus, we integrate these different LULC variables by making spatial predictions. We explicitly avoid making predictions on areas close to the river (lower right and upper left corners in Figure 20) because data were collected solely in tierra firme. Furthermore, we avoid extrapolating too much in covariate space by restricting predictions to areas where all the covariates are within the original range in the dataset  $\pm 10\%$ .

The spatial prediction for *A. darlingi* clearly reveals how both larva and mosquito biting rate tend to concentrate close to roads, particularly in areas with low forest cover and high amount of impervious soil (Figure 20).

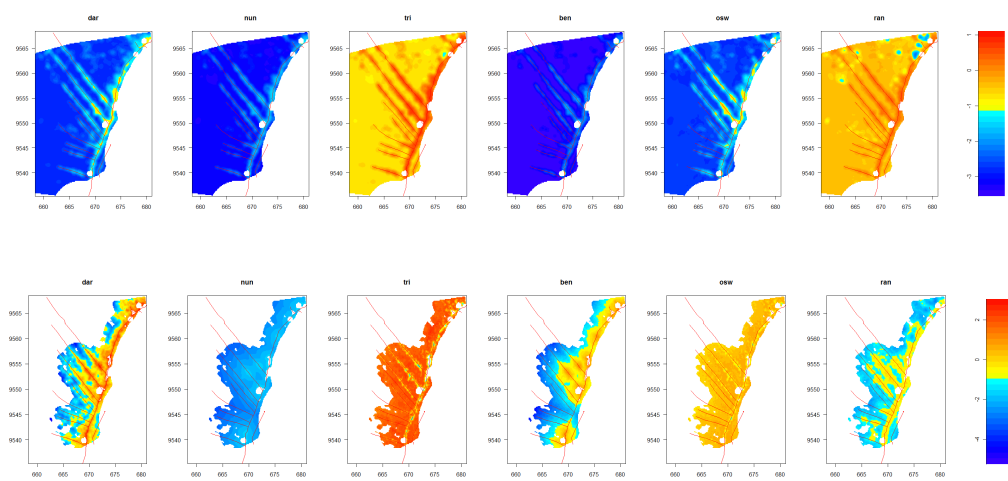


**Figure 20: Spatial prediction *A. darlingi* in relation to the number of water bodies with larva (left panel) and log mosquito biting rate (right panel).**

The network of roads is depicted as thin red lines.

However, a focus solely on *A. darlingi* ignores the factor that other vectors are far more abundant. As noted before, larva from *A. triannulatus* and *A. rangeli* are much more common than larva from *A. darlingi*. Although water bodies with larva from these species also concentrate around roads, they do not necessarily coincide with the areas where *A. darlingi* larva is abundant (upper panels in Figure 21). Similarly, spatial predictions for mosquito biting rate reveal the clear dichotomy between forested areas, where *A. triannulatus* and *A. oswaldoi* prevail, and anthropogenic areas, where *A. darlingi*

tends to dominate (lower panels in Figure 21). Furthermore, because the entire region is largely forested, *A. triannulatus* is by far the most common vector throughout this landscape. Finally, it is particularly striking how larvae from *A. triannulatus* and *A. oswaldoi* seem to be adapted to anthropogenic influence whereas adult mosquitoes from these same species thrive in more pristine areas.

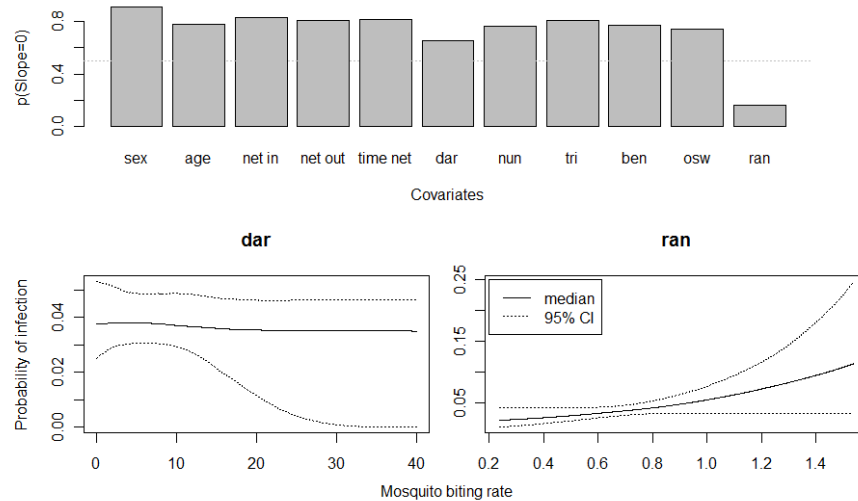


**Figure 21: Spatial prediction of the log number of water bodies with larva (upper panels) and log mosquito biting rate (lower panels) for anopheline species.**

The network of roads is depicted as thin red lines.

### 4.3.3 Malaria prevalence and mosquito biting rate

In relation to the malaria prevalence data, we find that the strongest predictor of infection status was the biting rate of *A. rangeli* while, surprisingly, there was no indication that the biting rate of *A. darlingi* was associated with infection risk (Figure 22). These results were robust to changing the time lag on predicted mosquito biting rate from two weeks to only one week.



**Figure 22: Predictors of malaria infection.**

Upper panel depicts the probability that the slope parameter  $\beta_4$  is equal to zero for each covariate (net in: the time that the individual enters the net, net out: the time that the individual exits the bed-net, time net: total number of hours the individual stays under the bed-net). A line was added at 0.5 for reference (dashed grey line). Lower panels depict the predicted relationship between probability of infection and mosquito biting rate for *A. darlingi* (lower left panel) and *A. rangeli* (lower right panel).

There are three possible interpretations for the results in Figure 22. One interpretation is that *A. rangeli* is the main vector responsible for malaria transmission within and in the proximities of these households. This could be the case if its abundance is higher than that of *A. darlingi* where and when observations were not made (i.e., observations of mosquito biting-rate were restricted to 18 – 24:00 and were made outside houses), a hypothesis strongly supported by the fact that *A. rangeli* was one of the most common larva species. The second hypothesis is that, despite being outnumbered by *A. darlingi* (notice the scale of the x-axis), *A. rangeli* is a far more

competent vectors (i.e., more likely to be infected and transmit the *Plasmodium*) than *A. darlingi*. We find this explanation to be implausible. Finally, an alternative interpretation is that there is a mismatch between our modeled mosquito biting rates, which are based on the household locations and a period from 18:00 to 24:00, and the location and time of the day that people are more exposed to being infected. If there is such a mismatch, then the associations we found might be spurious. For instance, malaria infection may be occurring away from where people live, perhaps when individuals engage in activities such as farming, logging, fishing, and hunting. In this case, predictions of mosquito biting rate based on household locations would be irrelevant. Alternatively, perhaps only a fraction of the observation period (i.e., 18-24:00) is relevant for malaria infection. Indeed, we find that 99.4% of the individuals in our sample used bed-nets 7 days a week and that 85% of the individuals entered their bed-nets at or before 20:00, suggesting that infections might be predominantly occurring before 20:00.

#### **4.4 Discussion**

Our study integrates data from two mosquito life stages (larva and adult mosquito) from multiple species and malaria prevalence, providing novel insights regarding the different pathways through which climate and LULC influence malaria prevalence in the Peruvian Amazon. We find striking patterns regarding changes in species composition as distance to forests changes, where *A. triannulatus* is the main

vector close to forests whereas *A. darlingi* is the main vector away from forests. Despite these findings, our results do not corroborate the hypothesis that deforestation increases malaria risk because, contrarily to the expected pattern, areas that were predicted to have higher *A. darlingi* biting rate did not have higher malaria prevalence.

Entomological studies can provide important insights regarding the role of climate and LULC on malaria risk. However, most entomological studies focus on a single mosquito population index (e.g., presence of larva (38, 118, 119, 121-126) or mosquito biting rate (31, 40, 149)) which are then assumed to directly influence malaria risk. This assumption may or may not be valid. For instance, although higher abundance of water bodies with larva increases the probability of finding the adult mosquito, our results suggest that areas with high number of water bodies with larva are not necessarily the ones with high mosquito biting rate from the same species (e.g., results for *A. oswaldoi* in Figure 21). Similarly, our findings greatly corroborate the earlier published results regarding the increase in *A. darlingi* larva presence and mosquito biting rate in deforested sites; yet, the integration of entomological and epidemiological data do not suggest that this *A. darlingi* increase results in enhanced malaria risk.

Entomological studies also typically focus on one (31, 38, 40) or a small (150, 151) subset of mosquito species that are deemed important for malaria transmission. However, the use of data from multiple species allowed us to identify important similarities and differences regarding how larvae and biting rate from different species

respond to LULC and climate covariates. Furthermore, comparing the abundance of the different malaria vector species allowed us to put our *A. darlingi* findings into perspective. For instance, while *A. darlingi* seems to be much more abundant in anthropogenic areas, overall mosquito biting rate was actually higher closer to forests (Figure 19). Furthermore, only by using data from multiple species were we able to detect the striking partitioning of the landscape, with some species (e.g., *A. triannulatus* and *A. oswaldoi*) being the dominant vector in less anthropogenic areas while other species (e.g., *A. darlingi*) thrived in areas more heavily influenced by humans.

Malaria prevalence studies (e.g., 22, 65, 141-148), on the other hand, have the potential to uncover important predictors of malaria risk but rarely do they inform why these predictors are important. For example, one may detect rain to be an important malaria risk factor but it might not be clear if precipitation affects the mosquito breeding habitat, adult mosquito survival, or even human behavior. Integrating vector and disease data is therefore important to generate a better understanding of the causal pathways regarding how environmental factors affect malaria incidence. This knowledge is critical to guide malaria prevention and control activities, suggesting where and when specific malaria prevention and control strategies targeting the larva (e.g., larvicide application), the adult mosquito (e.g., indoor residual spraying, and use of bed-nets) or humans (e.g., chemotherapy and educational campaigns) will be more effective.

Considerable controversy has spurred regarding the relationship between deforestation and malaria risk. We believe that there are at least three reasons for the many conflicting views that have been published regarding this relationship. First, part of the difficulty of comparing different studies lies on the different types of land use that are generically referred to as deforested sites. For instance, in the Brazilian Amazon, forests are often cleared to give way to pasture while in Peru the main deforestation driver is swidden-fallow agriculture and deforested areas are often covered by shrubs and secondary vegetation growth (38, 40). Unfortunately, Landsat imagery might be too coarse to distinguish between these land uses. Second, the scale of deforestation might also be important because the relationship between human biting rate and deforestation may be non-linear. For instance, the data we use here include very few households that live with no vegetation within a 1,000 m radius around their house (less than 0.1% of the households have less than 5% vegetation within this radius). On the other hand, data from a study that found that *A. darlingi* depended on forested sites (31) came from a much more deforested region, where 30% of the households lived in areas with 5% or less vegetation within the same radius. Unfortunately, it is often very hard to encompass such a large range of deforestation levels in a single study, requiring either a very large spatial scale or a relatively long-term study.

Finally, several factors can influence malaria transmission beyond mosquito biting rate. For instance, the interpretation of our spatial predictions of mosquito biting

rate (Figure 21) fundamentally depends on the relative competence of the different mosquito species. Similarly, our mosquito biting rate observations are based on the same level of exposure (i.e., number of mosquitoes landing on exposed legs between 18 - 24:00) at the different sites. However, individuals may have different levels of exposure depending on their location (e.g., at home or outdoors), the activity they are engaged in, and the duration and time of the day at which this happens. For instance, virtually all the individuals in our sample consistently slept under bed-nets, potentially having lower exposure indoors than when going fishing at night. Understanding where and when individuals are being infected is critical to appropriately determine the implication of LULC changes for malaria risk, for instance by indicating which mosquito species is the main responsible for malaria transmission (i.e., *A. triannulatus* may be the main malaria vector if people are predominantly infected in the forest). This finding suggests caution regarding a) using data collected solely in the vicinities of where people live to infer the main malaria vector species because of the implicit assumption that infection occurs within or close to houses; and b) inference on malaria risk based solely on mosquito data, thus ignoring how human behavior may alter the level of exposure. Indeed, a recent meta-analysis suggests that the relationship between mosquito density and malaria risk is far from simple (152).

Our study has four important limitations. First, the malaria prevalence data had a very low prevalence of infection (as detected by microscopy) (~4%), which precluded

the identification of multiple risk factors. We believe this low infection prevalence might be due to the detection method employed; it is widely acknowledged that a more sensitive diagnostic method than microscopy (e.g., PCR) is critical in settings where parasitemia levels are low and a large pool of individuals with sub-patent malaria infections may exist (17). Second, some of the secondary mosquito species might actually be species complexes, with the individual species potentially having distinct behavioral and habitat characteristics (e.g., *A. triannulatus*; 153, 154). Third, an important factor that we have not considered in our analysis refers to the differential susceptibility to infection of the mosquito species we analyzed (155). Fourth, although we modeled each dataset separately, a better approach would have been to make a more coherent analysis by specifying a joint model, this way fully propagating the uncertainty on the individual parameters. Unfortunately this approach would have substantially increased the complexity of our analysis, particularly in light of the multiple models we have evaluated (i.e., for different covariate sets). We believe that developing such a joint model remains an important area of research.

In summary, our study highlights how climate and LULC strongly influence larva presence, whereas adult mosquito survival tended to be influenced mostly by LULC covariates. We describe striking patterns regarding how species composition of mosquito biting rate changes as a function of distance to forest cover, large water bodies, and impervious areas. Finally, our results strongly corroborate earlier findings

regarding an increase in *A. darlingi* abundance in deforested sites but we did not find an association between this increase in *A. darlingi* biting rate and malaria prevalence. We believe that this lack of association might be due to a mismatch between the predicted mosquito biting rates, which assume that infection from 18:00 to 24:00 outside but close to households, and the location and time of the day in which higher exposure to infection really occurs. Our article illustrates the novel insights that can be gained through the integration of entomological data from multiple species and mosquito life stages with malaria prevalence data.

## Appendix I

### 1. *Description of covariates*

Whenever linking secondary environmental data to malaria data, it is crucial to determine the area surrounding each household that can be reasonably expected to influence the observed *Plasmodium* infection risk (onwards 'area of influence'). For our study, we set the area of influence for each household as a circle with radius of 1,000m centered at the household. We chose this radius because it reflects the area where we expect people to spend most of their time and because the malaria vector is known to have a flight range on this order of magnitude (see review in 66, 77, 156).

The covariates for infection risk were:

- *gender* (binary, women=1 and men=0);
- *chain sawyer* (binary, yes=1 and no=0): if works primarily as a chain sawyer;
- *age* (in years);
- *education* (in years of schooling);
- *time in Acrelandia* (in years): Clinical and parasitological immunity is often higher with greater past exposure to malaria (e.g., 157, 158). Because people in rural land settlements often come from malaria-free regions (22), we added the time living in Acrelandia as a proxy for past exposure to malaria;
- *extractivism* (binary, yes=1 and no=0): if participates in extractivism activities;

- *hunt/fish* (binary, yes=1 and no=0): if participates in hunting or fishing activities;
- *co-inhabits*  $D^m=1$  (binary, yes=1 and no=0): if the person being tested shares their house with somebody that had a positive microscopy result in the past 30 days;
- *co-inhabits*  $D^{pcr}=1$  (binary, yes=1 and no=0): if the person being tested shares their house with somebody that had a positive PCR result in the past 30 days;
- *water area* (surface water area in 2008, in ha): this was estimated by visual interpretation of high-spatial resolution imagery (2 × 2m and 8 × 8m pixels for panchromatic and multispectral images, respectively), acquired by FORMOSAT in 2008. Many of these water bodies can be considered permanent landscape features (e.g., natural rivers, ponds created to raise fish or provide water to the cattle). Thus, in the absence of high-spatial resolution imagery from earlier years, we assumed that these water bodies were probably present throughout 2004-2008.
- *forest area* (in ha in 2004): the Brazilian Space Agency (INPE) provides yearly land cover classification maps for the entire Brazilian Amazon based on a semi-automated analysis of Landsat imagery (84). Using these maps, we were able to determine forest extent in 2004 within the area of influence of each household;

- *deforestation rate* (in ha yr<sup>-1</sup>): using data from INPE (described above), we determined the yearly deforested area within the area of influence of each household;
- *precipitation* (monthly average, in mm hour<sup>-1</sup>): Precipitation data came from the Tropical Rainfall Measuring Mission - TRMM (85). We used the '3B43 Monthly 0.25 x 0.25 degree merged TRMM and other sources estimates' product, with a one month time lag. The assumption regarding this time lag is that water affects the vector mainly through its breeding habitat. Therefore, changes in precipitation should only affect infection risk on the following month since this is the minimum necessary time for the larvae to become an adult, the adult to be infected and finally the adult to become infectious.
- *drought index* (monthly, in mm): this drought index is determined by an algorithm that takes into account precipitation and evapotranspiration to calculate water deficit. The details of this algorithm are given in Aragao *et al.* (159). Despite its simplicity, this index has been used extensively to characterize drought in the region (87, 88, 159). Precipitation data came from TRMM (85) and we used a one month time lag with the same rationale as for precipitation.

We also added interaction terms involving precipitation/forest area, water deficit/forest area and water area/forest area. All continuous covariates were

standardized by subtracting their mean and dividing by their standard deviation. None of the covariates listed above were highly correlated (i.e.,  $|r| < 0.8$ )

## 2. Derivation of the likelihood

When only microscopy results are available for AACD, the likelihood is:

$$p(D^m) = \sum_{I=0,1} p(D^m | I) p(I) = \left[ \sum_{S=0,1} p(D^m | S, I=1) p(S | I=1) \right] p(I=1) + p(D^m | I=0) p(I=0)$$

When microscopy results and symptom statuses are available for AACD, the likelihood is:

$$p(D^m, S) = \sum_{I=0,1} p(D^m, S | I) p(I) = p(D^m | S, I=1) p(S | I=1) p(I=1) + p(D^m | I=0) p(S | I=0) p(I=0)$$

When only PCR results are available for AACD, the likelihood is:

$$p(D^{pcr}) = \sum_{I=0,1} p(D^{pcr} | I) p(I)$$

When PCR results and symptom statuses are available for AACD, the likelihood is:

$$p(D^{pcr}, S) = \sum_{I=0,1} p(D^{pcr}, S | I) p(I) = \sum_{I=0,1} p(D^{pcr} | I) p(S | I) p(I)$$

When PCR and microscopy results are available for AACD, the likelihood is:

$$\begin{aligned} p(D^{pcr}, D^m) &= \sum_{I=0,1} p(D^{pcr}, D^m | I) p(I) = \sum_{I=0,1} p(D^{pcr} | I) p(D^m | I) p(I) \\ &= p(D^{pcr} | I=1) \left[ \sum_{S=0,1} p(D^m | S, I=1) p(S | I=1) \right] p(I=1) + p(D^{pcr} | I=0) p(D^m | I=0) p(I=0) \end{aligned}$$

When symptom statuses, PCR and microscopy results are available for AACD, the likelihood is:

$$p(D^{pcr}, D^m, S) = \sum_{I=0,1} p(D^{pcr}, D^m, S | I) p(I) = \sum_{I=0,1} p(D^{pcr}, D^m | S, I) p(S | I) p(I) = \sum_{I=0,1} p(D^{pcr} | I) p(D^m | S, I) p(S | I) p(I)$$

When only microscopy results are available for ACD (or PCD, if we substitute

ACD for PCD in the formulae below), the likelihood is:

$$\begin{aligned} p(D^m | ACD) &= \sum_{S=0,1} p(D^m | S, ACD) p(S | ACD) = \sum_{S=0,1} [\sum_{I=0,1} p(D^m | I, S, ACD) p(I | S, ACD)] p(S | ACD) \\ &= \sum_{S=0,1} [\sum_{I=0,1} p(D^m | I, S) p(I | S)] p(S | ACD) = \sum_{S=0,1} [\sum_{I=0,1} p(D^m | I, S) \frac{p(S | I) p(I)}{p(S)}] p(S | ACD) \end{aligned}$$

When microscopy results and symptom statuses are available for ACD (or PCD,

if we substitute ACD for PCD in the formulae below), the likelihood is:

$$\begin{aligned} p(D^m, S | ACD) &= \sum_{I=0,1} p(D^m, S | I, ACD) p(I | ACD) = \sum_{I=0,1} p(D^m | S, I, ACD) p(S | I, ACD) p(I | ACD) \\ &= \sum_{I=0,1} p(D^m | S, I) \frac{p(I | S, ACD) p(S, ACD)}{p(I, ACD)} p(I | ACD) = \sum_{I=0,1} p(D^m | S, I) p(I | S) p(S | ACD) \\ &= \sum_{I=0,1} p(D^m | S, I) p(S | I) p(I) \frac{p(ACD | S)}{p(ACD)} \end{aligned}$$

### 3. Likelihood formulae

To avoid clutter in our equations, we denote  $A = p(I_{i,t} = 1) = \frac{1}{1 + e^{-(x_{i,t}^T \beta + \phi_i + \theta_{h(i)})}}$ ,

$B = p(S_{i,t} = 1 | I_{i,t} = 1) = \frac{1}{1 + e^{-(x_{i,t}^T \gamma)}}$ , and  $C = p(S_{i,t} = 1 | I_{i,t} = 0)$ . For ACD,  $E$  and  $F$  denote

$p(ACD | S_{i,t} = 1)$  and  $p(ACD | S_{i,t} = 0)$ , respectively. For PCD,  $E$  and  $F$  denote  $p(PCD | S_{i,t} = 1)$

and  $p(PCD | S_{i,t} = 0)$ , respectively. We chose to suppress subscripts but it is understood

that everything is for a given individual  $i$  residing at household  $h$  at time  $t$ . The

likelihood formulae using the above notation are given in Table 9 and Table 10.

**Table 9: Likelihood of each of the possible outcomes in AACD**

$D^{pcr}$	$D^m$	S	Likelihood
X	0	X	$[(1-\alpha_1)B+(1-\alpha_0)(1-B)]A+(1-A)$
X	1	X	$[\alpha_1B+\alpha_0(1-B)]A$
0	X	X	$(1-\delta)A+\pi(1-A)$
1	X	X	$\delta A+(1-\pi)(1-A)$
X	1	0	$\alpha_0(1-B)A$
X	0	0	$(1-\alpha_0)(1-B)A+(1-C)(1-A)$
X	0	1	$(1-\alpha_1)BA+C(1-A)$
X	1	1	$\alpha_1BA$
1	X	0	$\delta(1-B)A+(1-\pi)(1-C)(1-A)$
0	X	0	$(1-\delta)(1-B)A+\pi(1-C)(1-A)$
0	X	1	$(1-\delta)BA+\pi C(1-A)$
1	X	1	$\delta BA+(1-\pi)C(1-A)$
1	0	X	$\delta[(1-\alpha_1)B+(1-\alpha_0)(1-B)]A+(1-\pi)(1-A)$
0	0	X	$(1-\delta)[(1-\alpha_1)B+(1-\alpha_0)(1-B)]A+\pi(1-A)$
0	1	X	$(1-\delta)[\alpha_1B+\alpha_0(1-B)]A$
1	1	X	$\delta[\alpha_1B+\alpha_0(1-B)]A$
0	0	0	$(1-\delta)(1-\alpha_0)(1-B)A+\pi(1-C)(1-A)$
1	1	1	$\delta\alpha_1BA$
0	0	1	$(1-\delta)(1-\alpha_1)BA+\pi C(1-A)$
0	1	0	$(1-\delta)\alpha_0(1-B)A$
1	0	0	$\delta(1-\alpha_0)(1-B)A+(1-\pi)(1-C)(1-A)$
1	1	0	$\delta\alpha_0(1-B)A$
0	1	1	$(1-\delta)\alpha_1BA$
1	0	1	$\delta(1-\alpha_1)BA+(1-\pi)C(1-A)$

**Table 10: Likelihood of each of the possible outcomes in ACD and PCD.**

$D^{pcr}$	$D^m$	S	Likelihood
X	0	X	$\frac{[(1-\alpha_1)BE+(1-\alpha_0)(1-B)F]A+[CE+(1-C)F](1-A)}{E(BA+C(1-A))+F[1-(BA+C(1-A))]}$
X	1	X	$\frac{[\alpha_1BE+\alpha_0(1-B)F]A}{E(BA+C(1-A))+F[1-(BA+C(1-A))]}$
X	1	0	$\frac{\alpha_0(1-B)FA}{E(BA+C(1-A))+F[1-(BA+C(1-A))]}$
X	0	0	$\frac{F[(1-\alpha_0)(1-B)A+(1-C)(1-A)]}{E(BA+C(1-A))+F[1-(BA+C(1-A))]}$
X	0	1	$\frac{E[(1-\alpha_1)BA+C(1-A)]}{E(BA+C(1-A))+F[1-(BA+C(1-A))]}$
X	1	1	$\frac{\alpha_1BEA}{E(BA+C(1-A))+F[1-(BA+C(1-A))]}$

#### **4. Full conditional distribution for the parameters sampled via a Gibbs sampling step**

The parameters that were updated via a Gibbs sampling step were the variances for the individual and household-level random effects. The full conditional distributions for these parameters are given below:

$$\frac{1}{\sigma_{ind}^2} \sim \text{Gamma}\left(\frac{N_{ind} - 1}{2}, \frac{\sum_{i=1}^{N_{ind}} \varphi_i}{2}\right) I(0 < \sigma_{ind} < 100)$$

$$\frac{1}{\sigma_h^2} \sim \text{Gamma}\left(\frac{N_h - 1}{2}, \frac{\sum_{h=1}^{N_h} \mathcal{G}_h}{2}\right) I(0 < \sigma_h < 100)$$

where  $\varphi_i$  and  $\mathcal{G}_{h[i]}$  are individual and household random effects, respectively, and

$N_{ind}$  and  $N_{hh}$  are the total number of individuals and households, respectively.

#### **5. Description of how data were simulated**

To make a realistic comparison of the different methods, we tried to mimic the original dataset as closely as possible. Therefore, we used the same covariate values as in the original dataset to simulate the data. To evaluate how reliably each method estimated effects of different sizes (as well as no effect at all), we assigned one of the following values to the risk factor parameters, both of infection and symptoms given infection: -0.5, -0.2, -0.1, 0, 0.1, and 0.2. All the remaining parameters were assigned values close to what we had already estimated in previous runs of our model. All these parameter values are summarized in Table 11.

**Table 11: Summary of parameter values adopted for the simulated data.**

Parameter	Description	Values
$\alpha_1$	Microscopy sensitivity given $S=1$	0.6
$\alpha_0$	Microscopy sensitivity given $S=0$	0.4
$\delta$	PCR sensitivity	0.8
$\pi$	PCR specificity	0.985
$\beta$	Covariates of infection risk factors	One of the following: -0.2, -0.1, 0, 0.1, and 0.2
$\sigma_{ind}$	Standard deviation of the individual level random effects	0.25
$\sigma_h$	Standard deviation of the household level random effects	0.5
$\gamma$	Covariates of risk factors of symptoms given infection	One of the following: 2, -0.5, 0, and 0.1
$p(S=1 I=0)$	Probability of symptoms given no infection	0.03
$p(PCD S=0)$	Probability of being sampled through PCD given no symptoms	0.012
$p(PCD S=1)$	Probability of being sampled through PCD given symptoms	0.6
$p(ACD S=0)$	Probability of being sampled through ACD given no symptoms	0.02
$p(ACD S=1)$	Probability of being sampled through ACD given symptoms	0.3

The simulated dataset had approximately the same number of microscopy and PCR results for the different sampling designs (AACD, PCD, and ACD) as the original dataset (Table 12).

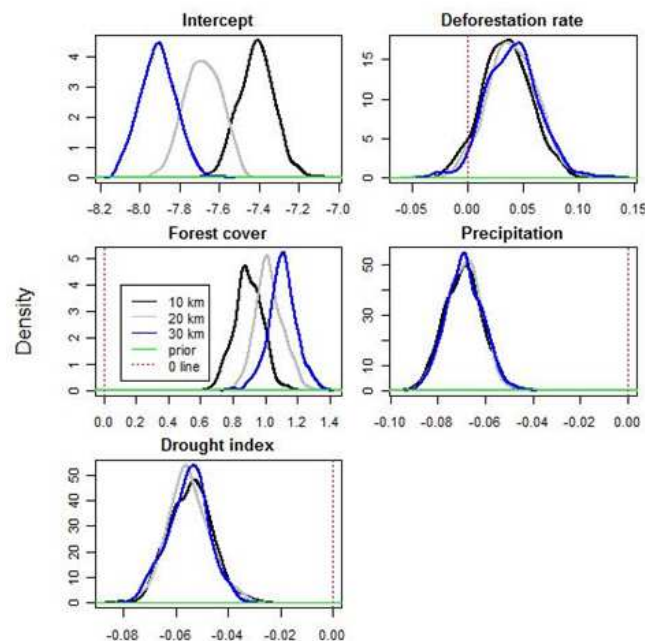
**Table 12: Number of microscopy and PCR results for the different sampling designs, both from the original and simulated datasets.**

Sampling design	Original dataset		Simulated dataset	
	PCR	Microscopy	PCR	Microscopy
AACD	1400	1383	1291	1291
ACD	0	940	0	773
PCD	0	754	0	1294

## Appendix II

### 1. Comparison of modeling results using different radii for the catchment area.

We evaluated how robust our results were to the definition of catchment area by quantifying our covariates and population size  $N_i$  using different catchment area radii (i.e., 10, 20, and 30 km) and re-fitting our model. We find that the parameters that change the most were  $\beta_0$  (intercept) and  $\beta_2$  (forest cover effect) (Figure 23). However, these changes do not modify our overall conclusions, namely that the main determinant of malaria cases is forest cover, deforestation rate has a small but positive effect, and precipitation and drought index have a small but consistently negative effect.



**Figure 23: Posterior distribution of the main regression parameters with covariates and population size assessed using three different catchment area radii (10, 20, and 30 km).**

A line at zero (dashed red line) was added for reference.

## **2. Alternative model formulation.**

The model in the main text was fitted using the pooled data from seven Brazilian Amazon states. Another approach is to use city-specific regressions and pool information from all these models, essentially exploring the within-city (rather than within-city and between-city) variability. The model is specified as follows:

$$C_{iym} \sim \text{Poisson}(\exp(w_{iym})N_i)$$

where  $N_i$  is the population size within the catchment area, and  $w_{iym}$  is given by:

$$w_{iym} \sim N(\beta_{0i} + \beta_{1i}X_{iy}, \sigma^2)$$

We fitted two models, one where  $N_i$  was deforestation rate and one where  $X_{iy}$  was forest cover. This model specification allows for city specific intercepts  $\beta_{0i}$  and slopes  $\beta_{1i}$ . We can pool information from these parameters by assigning a random effects prior to them. Our priors were:

$$\beta_{0i} \sim N(\beta_0, \tau^2)$$

$$\beta_{1i} \sim N(\beta_1, \gamma^2)$$

$$\beta_0, \beta_1 \sim N(0, 100)$$

$$\sigma, \tau, \gamma \sim Unif(0, 100)$$

These statistical models revealed that the pooled forest cover slope was positive and significantly different from zero, and much larger than the pooled deforestation rate slope (Table 13).

**Table 13: Summary statistics of the posterior distribution of the pooled forest cover effect ( $\beta_1^{for}$ ) and deforestation rate effect ( $\beta_1^{def}$ ) for the alternative model.**

LCI and UCI: lower and upper limit of the 95% credible interval

Parameter	Mean	LCI	UCI
$\beta_1^{def}$	0.09	-0.02	0.20
$\beta_1^{for}$	5.65	1.68	9.63

### 3. Auxiliary figures and tables

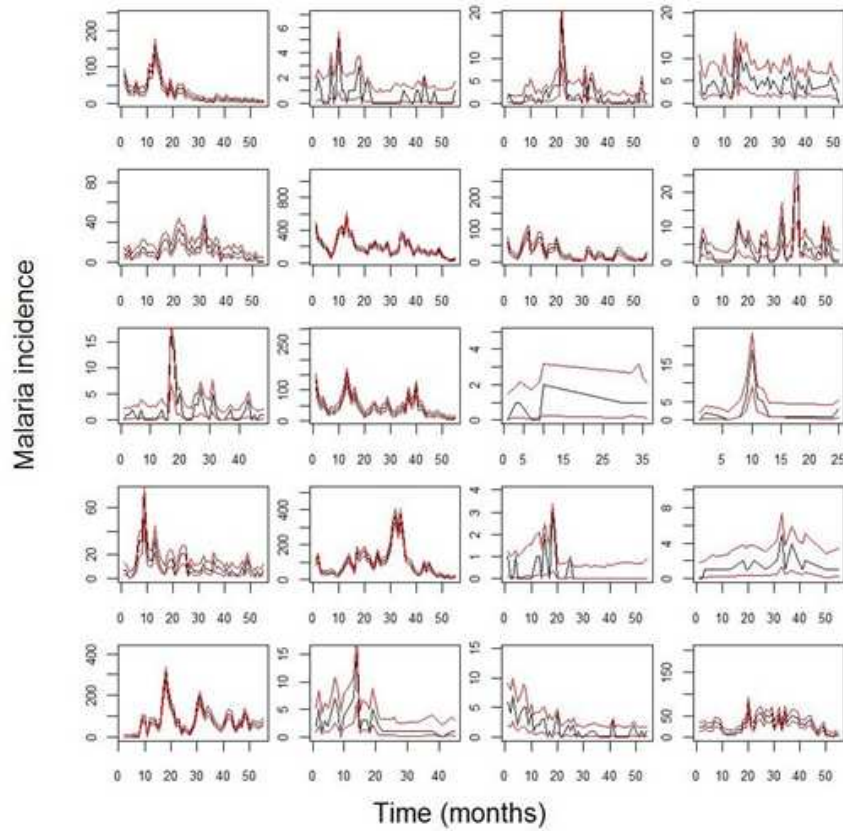
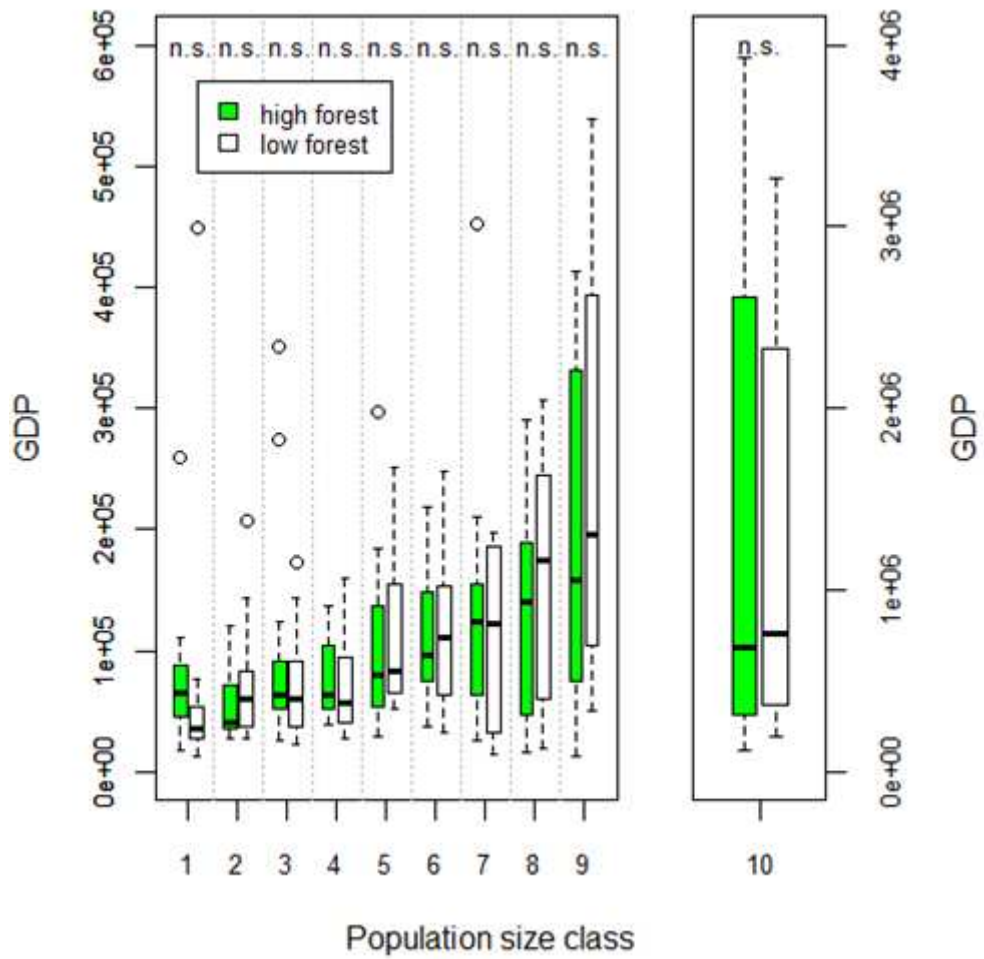


Figure 24: Comparison of the data (black line) and the 95% posterior predictive interval (red lines) for 20 randomly chosen cities.

Table 14: Convergence statistic  $R(91)$  for the regression parameters (intercept and slopes for the different covariates) in the main model.

Parameter	Point estimate	97.5% quantile
$\beta_0$	0.999	1.00
$\beta_1$	1.002	1.01
$\beta_2$	1.006	1.02
$\beta_3$	0.999	1.00
$\beta_4$	1.007	1.01



**Figure 25: Gross domestic product is similar in cities with low and high forest cover.**

Data were stratified into 10 percentile population size classes and average gross domestic product (GDP) for each year and city was depicted. Within each size class, we compare cities with high (green box-plots) vs. low forest cover (white box-plots). Cities with high forest cover are cities that have forest cover higher than the median for that size class. 'n.s.', '\*', '\*\*', and '\*\*\*\*' are non-significant ( $p > 0.05$ ), significant ( $0.01 < p < 0.05$ ), very significant ( $0.001 < p < 0.01$ ) and highly significant ( $p < 0.001$ ) difference in means, respectively, based on permutation tests.

## Appendix III

### 1. Simulation results

Our simulation assumes that we have 25 sites in each area and that these areas are either forested or deforested. In each site  $i$  ( $i=1, \dots, 25$ ), we sample water bodies using one transect. We further assume that the forested area has an average of 30 water bodies per transect ( $E[W_i^{for}] = \lambda^{for} = 30$ ), from which an average 27% have larva ( $p^{for} = 0.2\bar{6}$ ), while the deforested area tends to have less water bodies on average per transect ( $E[W_i^{def}] = \lambda^{def} = 10$ ) but a greater proportion of them tend to have larva ( $p^{def} = 0.8$ ).

Let the number of water bodies with larva per transect for site  $i$  be denoted as  $L_i^{for}$  and  $L_i^{def}$ , for the forested and deforested sites, respectively. In our simulations, we assume that:

$$W_i^{for} | \lambda^{for} \sim \text{Poisson}(\lambda^{for})$$

$$W_i^{def} | \lambda^{def} \sim \text{Poisson}(\lambda^{def})$$

$$L_i^{for} | W_i^{for}, p^{for} \sim \text{Binom}(W_i^{for}, p^{for})$$

$$L_i^{def} | W_i^{def}, p^{def} \sim \text{Binom}(W_i^{def}, p^{def})$$

With these assumptions, it is trivial to show that the expected number of water bodies with larva per transect is the same for the forested and deforested sites (i.e.,  $E[L_i^j] = E[E[L_i^j | W_i^j]] = E[W_i^j p^j] = \lambda^j p^j = 8$ , where  $j = def$  or  $j = for$ ). Using data generated in this fashion, we can run a binomial regression to show that there is a

significant difference between deforested and forested sites on *the proportion of water bodies with larva*, despite the same average *number of water bodies with larva per transect*.

Alternatively, we can consider the abundance of larva per transect instead of the number of water bodies with larva per transect. We now assume that we have only two sites, one forested and the other deforested. We further assume that the forested site has an average larva abundance per water body of 3.3 ( $\delta^{for} = 3.\bar{3}$ ) whereas the deforested area has a higher average larva abundance per water body ( $\delta^{def} = 10$ ). Let the number of larvae for water body  $i$  be denoted as  $L_i^{for}$  and  $L_i^{def}$ , for the forested and deforested sites, respectively. We now assume that:

$$W^{for} | \lambda^{for} \sim Poisson(\lambda^{for})$$

$$W^{def} | \lambda^{def} \sim Poisson(\lambda^{def})$$

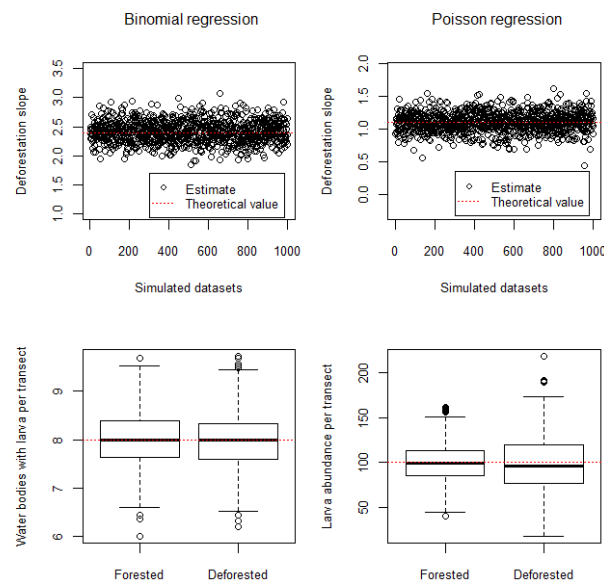
$$L_i^{for} | \delta^{for} \sim Poisson(\delta^{for}) \text{ for each water body}$$

$$L_i^{def} | \delta^{def} \sim Poisson(\delta^{def}) \text{ for each water body}$$

Again, the expected number of larva per transect is the same for the forested and deforested site (i.e.,  $E[\sum_{i=1}^{W^j} L_i^j] = E[E[\sum_{i=1}^{W^j} L_i^j | W^j]] = E[W^j \delta^j] = \lambda^j \delta^j = 100$ , where  $j=def$  or

$j=for$ ). Using these data, we can run a Poisson regression to show that there is a significant difference between the forested and deforested site in terms of the *larva abundance per water body*, despite the same average *larva abundance per transect*.

We performed 1000 simulations and show that both the binomial and the Poisson regressions consistently estimate a statistically significant difference between forested and deforested sites (upper panels in Figure 26), despite the fact that these datasets were created so that both sites have the same average number of water bodies with larva or larva abundance per transect (lower panels in Figure 26).



**Figure 26: Regression models detect a significant difference between forested and deforested sites (upper panels), despite the same average per transect (lower panels).**

Simulation results show that Binomial (left panels) and Poisson (right panels) regressions can indicate a significant difference between the forested and deforested sites (upper panels) even if the average number of water bodies with larva or larva abundance per transect is the same for both sites (lower panels).

## **2. Description of the binomial model used to fit the *A. darlingi* larva data.**

### **2.1. Data**

We surveyed larval anophelines in 56 locations (i.e., 14 sites x 4 vegetation types within each site). Larvae were collected once every three weeks between March and August 2001 (resulting in 8 collections) from water bodies found along multiple transects. Further details regarding the larval collection methodology can be found in (38, 139).

Larva abundance data greatly depends on methodological details (e.g., where and how samples were taken as well as how many larvae died before being identified), resulting in considerable variability. To avoid this variability, we converted these data to presence/absence data. Data are summarized for each location  $j$  ( $j=1, \dots, 56$ ) and collection  $t$  ( $t=1, \dots, 8$ ) as the number of water bodies with larva  $L_{jt}$  and the total number of sampled water bodies  $W_{jt}$ .

### **2.2. Covariates**

The covariates we used were land use / land cover (LULC) and climate related covariates, estimated using remote sensing. The set of LULC variables includes urban, water, forest, and non-forest vegetation areas, estimated using an unsupervised classification of a 2001 Landsat image within a radius of 1000 m from the center of each larva transect.

The climate covariates include precipitation and solar radiation at the day of collection and average (over the last 5 days prior to the collection date) minimum temperature, solar radiation and soil moisture. Solar radiation and minimum daily temperature were drawn from the Global Data Assimilation System (GDAS; 160, 0.471 degree resolution) gridded analysis and were topographically downscaled to 1km resolution using standard lapse rate corrections (e.g., 161). Precipitation estimates are from the Tropical Rainfall Measurement Mission (TRMM) Multisensor Precipitation Analysis (TMPA; 162). The three hourly, 25km gauge corrected estimates were used (product 3B42v7). GDAS meteorological fields and TRMM precipitation were applied as forcing data to offline simulations with the Noah Land Surface Model v3.2 (163, 164), implemented in the NASA Land Information System (165). Noah simulations were used to generate estimates of surface and root zone soil moisture every three hours over the period of analysis.

### 2.3. Methods

We employ a binomial model

$$L_{jt} | W_{jt}, p_{jt} \sim \text{Binomial}(W_{jt}, p_{jt}),$$

where  $p_{jt}$  is assumed to be a function of our covariates

$$\text{logit}(p_{jt}) = \alpha + \mathbf{x}_{jt}^T \boldsymbol{\beta} .$$

We adopt vague priors for the intercept  $\alpha$  but multiple shrinkage prior for the slope parameters  $\beta$ , which allows for priors to be stronger or weaker depending on the magnitude of the slope parameter:

$$\alpha \sim N(0,10) ,$$

$$\beta_k | \tau_k^2 \sim N(0, \tau_k^2),$$

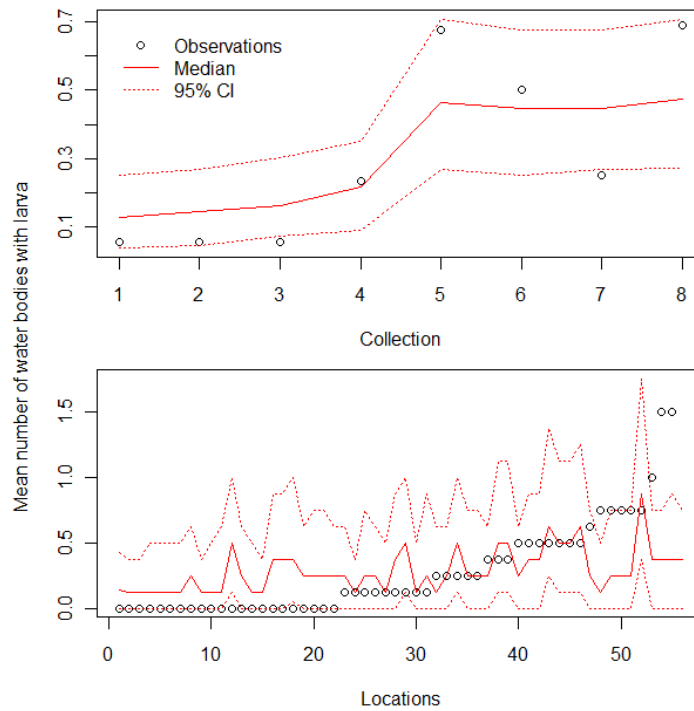
$$\tau_k^2 | \lambda \sim Exponential(\lambda) .$$

We chose  $E[\tau_k^2] = 1 / \lambda = 0.01$  which implies  $\lambda = 100$ . This assumes that the prior on the slope parameter will have approximately an average standard deviation of 0.1.

This model was fitted using a Gibbs sampler, with Metropolis-within-Gibbs steps to sample the full conditionals. On total, 100,000 iterations were run and the first 10,000 iterations were discarded as burn-in. Convergence was assessed through trace-plots.

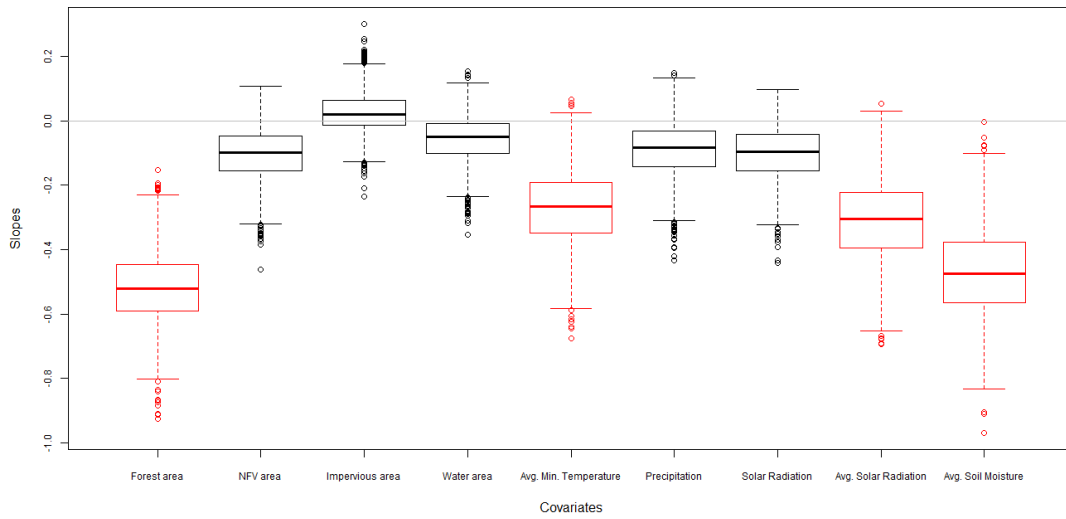
## 2.4. Results

The comparison of the posterior predictive distribution and the larva data suggests an adequate fit of our model (Figure 27).



**Figure 27: Comparison of the predicted and observed average number of water bodies with *Anopheles darlingi* larva, for different collections periods (upper panel) and locations (lower panel).**

The posterior distribution of the slope parameters suggest that forest area and average minimum temperature, solar radiation and soil moisture, significantly influence the probability of finding *A. darlingi* larvae in water bodies (Figure 28).



**Figure 28: Posterior distribution of slope parameters.**

Red boxes indicate slopes that are significantly different from zero. NFV stands for non-forest vegetation area and Avg. denotes average for 5-days prior to the collection date.

## Appendix IV

### 1. Priors for the larva model

In this model, we assume vague priors for the intercept and zero-inflation probability

$$\alpha_{1s} \sim N(0,10),$$

$$\omega_s \sim Unif(0,1).$$

On the other hand, we impose a multiple shrinkage prior on the slope parameters, which allows for priors to be stronger or weaker depending on the magnitude of these slope parameter, by assuming that

$$\boldsymbol{\beta}_{1s} | \mathbf{T}_1^2 \sim N(\mathbf{0}, \mathbf{T}_1^2),$$

where  $\mathbf{T}_1^2$  is a diagonal matrix with elements  $\tau_{11s}^2, \dots, \tau_{1Ks}^2$ . We originally used a more conventional t prior but preliminary analysis revealed that this type of prior imposed very little shrinkage. As a result, we decided to adopt an exponential distribution for the variance parameter. We chose  $E[\tau_{1ks}^2] = 1/\lambda = 0.01$  which implies  $\lambda = 100$ . This assumes that the prior on the slope parameter will have approximately an average standard deviation of 0.1. More succinctly

$$\tau_{11s}^2, \dots, \tau_{1Ks}^2 \sim Exponential(100).$$

## 2. Priors for the biting rate model

In this model, we assume vague priors for the zero-inflation intercept and slope parameters and the intercept parameter from the mean of the negative binomial distribution

$$\beta_{2s}, \alpha_{2s}, \alpha_{3s} \sim N(0, 10).$$

Similar to the larva model, we adopt a multiple shrinkage prior on the slope parameters

$$\boldsymbol{\beta}_{3s} | \mathbf{T}_3^2 \sim N(\mathbf{0}, \mathbf{T}_3^2),$$

where the diagonal matrix  $\mathbf{T}_3^2$  has entries  $\tau_{31s}^2, \dots, \tau_{3Ks}^2$ . Finally, we assign the following hyper-prior for these diagonal elements

$$\tau_{31s}^2, \dots, \tau_{3Ks}^2 \sim \text{Exponential}(100).$$

## 3. Latent state representation and priors for the probit regression used to model malaria prevalence

A probit regression can be specified as we have done in the main text

$$y_i \sim \text{Bernoulli}(\Phi(\alpha_4 + \mathbf{x}_i^T \boldsymbol{\beta}_4)).$$

Equivalently, one can adopt a latent state representation. To do this, let the latent state be given by

$$z_i = \alpha_4 + \mathbf{x}_i^T \boldsymbol{\beta}_4 + e_i,$$

where  $e_i \sim N(0, 1)$ . Then, we assume that  $y_i = 1$  if and only if  $z_i > 0$ .

Finally, we assume conjugate priors for the intercept  $\alpha_4$  and slope parameters

$\beta_4$ :

$$p(\alpha_4, \beta_4, \nu) = N\left(\begin{bmatrix} \alpha_4 \\ \beta_4 \end{bmatrix} \mid \mathbf{0}, \nu \begin{bmatrix} 1000 & 0 \\ 0 & \mathbf{I} \end{bmatrix}\right) IG(\nu \mid a, b).$$

#### 4. Model fit

Here we describe how well our larva and biting-rate models fitted the data by comparing the posterior predictive distribution to temporal averages. Overall, we find that these models adequately represented the uncertainty in our datasets (Figure 29 and Figure 30). For instance, we find that our 95% predictive credible intervals for the larva and mosquito biting-rate models included the observations 96% and 95% of the times, respectively.

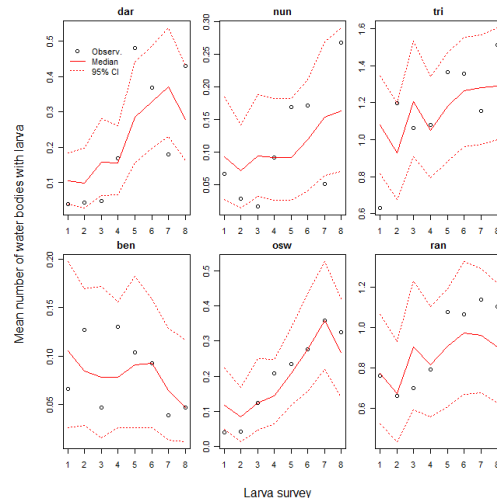
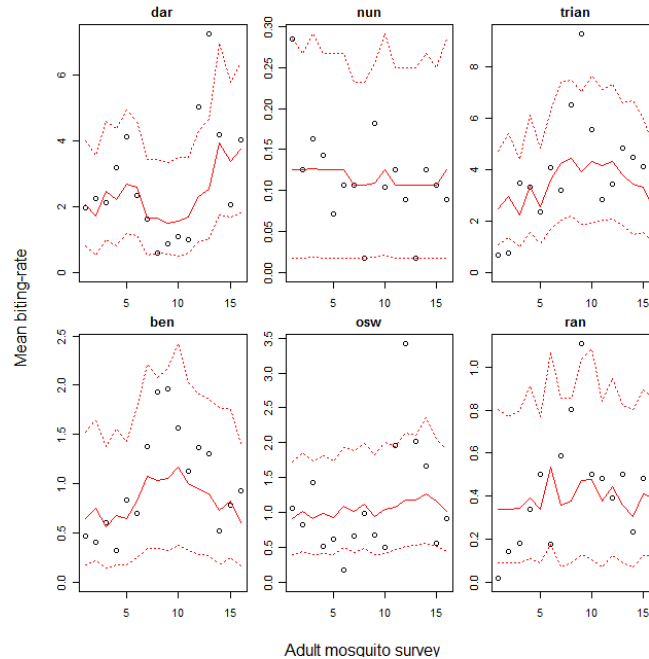


Figure 29: Comparison of the temporal trend on larva data (black circles) and the posterior predictive distribution (red lines).

Each panel shows the results for one of the six anopheline species.



**Figure 30: Comparison of the temporal trend on mosquito biting rate data (black circles) and the posterior predictive distribution (red lines).**

Each panel shows the results for one of the six anopheline species.

## **5. Description of climate covariates**

Solar radiation and minimum daily temperature were drawn from the Global Data Assimilation System (GDAS; 160, 0.471 degree resolution) gridded analysis and were topographically downscaled to 1 km resolution using standard lapse rate corrections (e.g., 161). Precipitation estimates are from the Tropical Rainfall Measurement Mission (TRMM) Multisensor Precipitation Analysis (TMPA; 162). The three hourly, 25 km gauge corrected estimates were used (product 3B42v7). GDAS meteorological fields and TRMM precipitation were applied as forcing data to offline simulations with the Noah

Land Surface Model v3.2 (163, 164), implemented in the NASA Land Information System (165). Noah simulations were used to generate estimates of surface and root zone soil moisture every three hours over the period of analysis.

## References.

1. R. W. Snow, C. A. Guerra, A. M. Noor, H. Y. Myint, S. I. Hay, The global distribution of clinical episodes of Plasmodium falciparum malaria. *Nature* **434**, 214 (2005).
2. F. P. Hardnett, R. M. Hoekstra, M. Kennedy, L. Charles, F. J. Angulo, Epidemiologic issues in study design and data analysis related to FoodNet activities. *Clin Infect Dis* **38**, S121 (2004).
3. A. Pruss-Ustun, C. Corvalan, *Preventing Disease through Healthy Environments. Towards an Estimate of the Environmental Burden of Disease.* (World Health Organization, Geneva, Switzerland, 2006). Available at: [http://www.who.int/quantifying\\_ehimpacts/publications/preventingdisease/en/](http://www.who.int/quantifying_ehimpacts/publications/preventingdisease/en/)
4. S. I. Hay, C. A. Guerra, A. J. Tatem, A. M. Noor, R. W. Snow, The global distribution and population at risk of malaria: past, present, and future. *Lancet Infect Dis* **4**, 327 (2004).
5. J. Sachs, P. Malaney, The economic and social burden of malaria. *Nature* **415**, 680 (2002).
6. N. Ravishankar *et al.*, Financing of global health: tracking development assistance for health from 1990 to 2007. *Lancet* **373**, 2113 (2009).
7. J. Oliveira-Ferreira *et al.*, Malaria in Brazil: an overview. *Malar J* **9**, 115 (2010).
8. M. C. Castro, R. L. Monte-Mor, D. O. Sawyer, B. H. Singer, Malaria risk on the Amazon frontier. *Proc Natl Acad Sci* **103**, 2452 (2006).
9. L. M. A. Camargo *et al.*, Hypoendemic malaria in Rondonia (Brazil, western Amazon region): seasonal variation and risk groups in an urban locality. *Am. J. Trop. Med. Hyg.* **55**, 32 (1996).

10. K. Laneri *et al.*, Forcing versus feedback: epidemic malaria and monsoon rains in Northwest India. *PLOS Computational Biology* **6**, (2010).
11. M. da Silva-Nunes, M. U. Ferreira, Clinical spectrum of uncomplicated malaria in semi-immune Amazonians: beyond the "symptomatic" vs "asymptomatic" dichotomy. *Mem. Inst. Oswaldo Cruz* **102**, 341 (2007).
12. M. U. Ferreira, M. Silva-Nunes, Evidence-based public health and prospects for malaria control in Brazil. *J Infect Dev Ctries* **4**, 533 (2010).
13. S. Ladeia-Andrade, M. U. Ferreira, M. E. Carvalho, I. Curado, J. R. Coura, Age-dependent acquisition of protective immunity to malaria in riverine populations of the Amazon Basin of Brazil. *Am. J. Trop. Med. Hyg.* **80**, 452 (2009).
14. C. Macauley, Aggressive active case detection: a malaria control strategy based on the Brazilian model. *Social Science & Medicine* **60**, 563 (2005).
15. FUNASA, *Manual de Terapeutica da Malaria*. (Ministerio da Saude. Superintendencia de Campanhas de Saude Publica - SUCAM., Brasilia, Brasil, 2001). Available
16. N. Silva *et al.*, Epidemiology and control of frontier malaria in Brazil: lessons from community-based studies in rural Amazonia. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **104**, 343 (2010).
17. L. C. Okell, A. C. Ghani, E. Lyons, C. J. Drakeley, Submicroscopic infection in *Plasmodium falciparum*-endemic populations: a systematic review and meta-analysis. *J. Infect. Dis.* **200**, 1509 (2009).
18. W. P. O'Meara, W. E. Collins, F. E. McKenzie, Parasite prevalence: a static measure of dynamic infections. *Am. J. Trop. Med. Hyg.* **77**, 246 (2007).
19. T. Hanscheid, M. P. Grobusch, How useful is PCR in the diagnosis of malaria? *Trends Parasitol.* **18**, 395 (2002).

20. A. Moody, Rapid diagnostic tests for malaria parasites. *Clinical Microbiology Reviews* **15**, 66 (2002).
21. M. da Silva-Nunes *et al.*, The Acre project: the epidemiology of malaria and arthropod-borne virus infections in a rural Amazonian population. *Cad Saude Publica* **22**, 1325 (2006).
22. M. da Silva-Nunes *et al.*, Malaria on the Amazonian frontier: transmission dynamics, risk factors, spatial distribution, and prospects for control. *Am. J. Trop. Med. Hyg.* **79**, 624 (2008).
23. D. L. Doolan, C. Dobano, J. K. Baird, Acquired immunity to malaria. *Clinical microbiology reviews* **22**, 13 (2009).
24. A. J. Branscum, I. A. Gardner, W. O. Johnson, Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. *Preventive Veterinary Medicine* **68**, 145 (2005).
25. C. Enoe, M. P. Georgiadis, W. O. Johnson, Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Preventive Veterinary Medicine* **45**, 61 (2000).
26. R. H. Barker Jr. *et al.*, Plasmodium falciparum and P. vivax: factors affecting sensitivity and specificity of PCR-based diagnosis of malaria. *Experimental Parasitology* **79**, 41 (1994).
27. K. F. Laserson *et al.*, Use of the polymerase chain reaction to directly detect malaria parasites in blood samples from the Venezuelan Amazon. *Am. J. Trop. Med. Hyg.* **50**, 169 (1994).
28. J. S. Clark, M. H. Hersh, Inference in incidence, infection, and impact: co-infection of multiple hosts by multiple pathogens. *Bayesian Anal* **4**, 337 (2009).

29. R Development Core Team, *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, Vienna, Austria, 2010). Available at: <http://www.R-project.org>
30. B. L. Hedt, M. Pagano, Health indicators: eliminating bias from convenience sampling estimators. *Statistics in Medicine* **30**, 560 (2010).
31. P. R. Moutinho, L. H. S. Gil, R. B. Cruz, P. E. M. Ribolla, Population dynamics, structure and behaviour of *Anopheles darlingi* in a rural settlement in the Amazon rainforest of Acre, Brazil. *Malar J* **10**, (2011).
32. S. H. Olson *et al.*, Links between climate, malaria, and wetlands in the Amazon basin. *Emerg Infect Dis* **15**, 659 (2009).
33. S. H. Olson, R. Gangnon, G. A. Silveira, J. A. Patz, Deforestation and malaria in Mancio Lima county, Brazil. *Emerg Infect Dis* **16**, 1108 (2010).
34. F. P. Alves *et al.*, High prevalence of asymptomatic *Plasmodium vivax* and *Plasmodium falciparum* infections in native Amazonian populations. *Am. J. Trop. Med. Hyg.* **66**, 641 (2002).
35. K. K. G. Scopel, C. J. F. Fontes, A. C. Nunes, M. F. Horta, E. M. Braga, High prevalence of *Plasmodium malariae* infections in a Brazilian Amazon endemic area (Apiacas - Mato Grosso State) as detected by polymerase chain reaction. *Acta Tropica* **90**, 61 (2004).
36. T. H. Katsuragawa *et al.*, Malaria and hematological aspects among residents to be impacted by reservoirs for the Santo Antonio and Jirau Hydroelectric power stations, Rondonia State, Brazil. *Cad. Saude Publica* **25**, 1486 (2009).
37. M. C. Castro, B. Singer, in *XXIV General population conference*. (Salvador, Bahia, Brasil, 2001). Available

38. A. Y. Vittor *et al.*, Linking deforestation to malaria in the Amazon: characterization of the breeding habitat of the principal malaria vector, *Anopheles darlingi*. *Am. J. Trop. Med. Hyg.* **81**, 5 (2009).
39. E. C. Oliveira, E. S. Santos, P. Zeihofer, R. Souza-Santos, M. Atanaka-Santos, Spatial patterns of malaria in a land reform colonization project, Juruena municipality, Mato Grosso, Brazil. *Malar J* **10**, (2011).
40. A. Y. Vittor *et al.*, The effect of deforestation on the human-biting rate of *Anopheles darlingi*, the primary vector of *Falciparum malaria* in the Peruvian Amazon. *Am. J. Trop. Med. Hyg.* **74**, 3 (2006).
41. C. A. Guerra, R. W. Snow, S. I. Hay, A global assessment of closed forests, deforestation and malaria risk. *Ann Trop Med Parasitol* **100**, 189 (2006).
42. J. Langhorne, F. M. Ndungu, A.-M. Sponaas, K. Marsh, Immunity to malaria: more questions than answers. *Nature Immunology* **9**, 725 (2008).
43. S. L. Schwartz, A. E. Gelfand, M. L. Miranda, Joint Bayesian analysis of birthweight and censored gestational age using finite mixture models. *Statistics in Medicine* **29**, 1710 (2010).
44. P. Slasor, N. Laird, Joint models for efficient estimation in proportional hazards regression models. *Statistics in Medicine* **22**, 2137 (2003).
45. I. A. Gardner, H. Stryhn, P. Lind, M. T. Collins, Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. *Preventive Veterinary Medicine* **45**, 107 (2000).
46. A. J. Branscum, I. A. Gardner, W. O. Johnson, Bayesian modeling of animal- and herd-level prevalences. *Preventive Veterinary Medicine* **66**, 101 (2004).
47. L. Joseph, T. W. Gyorkos, L. Coupal, Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology* **141**, 263 (1995).

48. W. O. Johnson, J. L. Gastwirth, L. M. Pearson, Screening without a "gold standard": the Hui-Walter paradigm revisited. *American Journal of Epidemiology* **153**, 921 (2001).
49. C.-L. Su, I. A. Gardner, W. O. Johnson, Diagnostic test accuracy and prevalence inferences based on joint and sequential testing with finite population sampling. *Statistics in Medicine* **23**, 2237 (2004).
50. N. Speybroeck *et al.*, True versus apparent malaria infection prevalence: the contribution of a Bayesian approach. *PLoS ONE* **6**, (2011).
51. W. P. O'Meara *et al.*, Sources of variability in determining malaria parasite density by microscopy. *Am. J. Trop. Med. Hyg.* **73**, 593 (2005).
52. M. B. Markus, The hypnozoite concept, with particular reference to malaria. *Parasitol. Res.* **108**, 247 (2011).
53. J. J. Juliano, N. Gadalla, C. J. Sutherland, S. R. Meshnick, The perils of PCR: can we accurately 'correct' antimalarial trials? *Trends Parasitol.* **26**, 119 (2010).
54. N. Myers, R. A. Mittermeier, C. G. Mittermeier, G. A. B. Fonseca, J. Kent, Biodiversity hotspots for conservation priorities. *Nature* **403**, 853 (2000).
55. R. S. DeFries, T. Rudel, M. Uriarte, M. Hansen, Deforestation driven by urban population growth and agricultural trade in the twenty-first century. *Nat Geosci* **3**, 178 (2010).
56. R. Walker *et al.*, Protecting the Amazon with protected areas. *Proc Natl Acad Sci* **106**, 10582 (2009).
57. M. C. Hansen *et al.*, Humid tropical forest clearing from 2000 to 2005 quantified by using multitemporal and multiresolution remotely sensed data. *Proc Natl Acad Sci* **105**, 9439 (2008).

58. J. Tollefson, Paying to save the rainforests. *Nature* **460**, 936 (2009).
59. T. H. Ricketts *et al.*, Indigenous lands, protected areas, and slowing climate change. *PLoS Biol* **8**, (2010).
60. Roll Back Malaria Partners, *The Global Malaria Action Plan for a Malaria-Free World*. (World Health Organization, Geneva, 2008). Available at: [www.rbm.who.int/gmap/](http://www.rbm.who.int/gmap/)
61. C. J. L. Murray *et al.*, Global malaria mortality between 1980 and 2010: a systematic analysis. *Lancet* **379**, 413 (2012).
62. M. L. Barreto *et al.*, Successes and failures in the control of infectious diseases in Brazil: social and environmental context, policies, interventions, and research needs. *Lancet* **377**, 1877 (2011).
63. USAID, *The Amazon Malaria Initiative: Overview*. (2012). Available at: <http://www.usaidami.org/extras/AMIFactsheet1overview.pdf>. Accessed 2012 March 19
64. The Global Fund, *The Global Fund: to Fight AIDS, Tuberculosis and Malaria*. (2012). Available at: <http://www.theglobalfund.org/>. Accessed 2012 March 19
65. D. Valle, J. Clark, K. Zhao, Enhanced understanding of infectious diseases by fusing multiple datasets: a case study on malaria in the Western Brazilian Amazon region. *PLOS One* **6**, e27462 (2011).
66. C. S. Castro, D. O. Sawyer, B. H. Singer, Spatial patterns of malaria in the Amazon: implications for surveillance and targeted interventions. *Health Place* **13**, 368 (2007).
67. F. S. M. Barros, M. E. Arruda, H. C. Gurgel, N. A. Honorio, Spatial clustering and longitudinal variation of *Anopheles darlingi* (Diptera: Culicidae) larvae in a river of the Amazon: the importance of the forest fringe and of obstructions to flow in frontier malaria. *Bull Entomol Res* **101**, 643 (2011).

68. C. E. A. Coimbra, Human factors in the epidemiology of malaria in the Brazilian Amazon. *Hum Organ* **47**, (1988).
69. A. F. Barbieri, D. O. Sawyer, B. S. Soares-Filho, Population and land use effects on malaria prevalence in the southern Brazilian Amazon. *Hum Ecol* **33**, 847 (2005).
70. M. Silva-Nunes *et al.*, Amazonian malaria: asymptomatic human reservoirs, diagnostic challenges, environmentally driven changes in mosquito vector populations, and the mandate for sustainable control strategies. *Acta Trop* **121**, 281 (2012).
71. J. A. Patz *et al.*, Unhealthy landscapes: policy recommendations on land use change and infectious disease emergence. *Environ Health Perspect* **112**, 1092 (2004).
72. J. A. Foley *et al.*, Global consequences of land use. *Science* **309**, 570 (2005).
73. J. A. Foley *et al.*, Amazonia revealed: forest degradation and loss of ecosystem goods and services in the Amazon Basin. *Front Ecol Environ* **5**, 25 (2007).
74. F. Keesing *et al.*, Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature* **468**, 647 (2010).
75. M. J. Pongsiri *et al.*, Biodiversity loss affects global disease ecology. *Bioscience* **59**, 945 (2009).
76. A. Dobson *et al.*, Sacred cows and sympathetic squirrels: the importance of biological diversity to human health. *PLoS Med* **3**, e231 (2006).
77. J. Keiser, B. H. Singer, J. Utzinger, Reducing the burden of malaria in different eco-epidemiological settings with environmental management: a systematic review. *Lancet Infect Dis* **5**, 695 (2005).

78. B. Singer, M. C. Castro, Enhancement and suppression of malaria in the Amazon. *Am. J. Trop. Med. Hyg.* **74**, 1 (2006).
79. Ministerio da Saude, *SIVEP Malaria*. (2010). Available at: [http://dw.saude.gov.br/portal/page/portal/sivep\\_malaria?Ano\\_n=2008](http://dw.saude.gov.br/portal/page/portal/sivep_malaria?Ano_n=2008). Accessed 2010 March 10
80. Ministerio de Saude, *Guia de vigilancia epidemiologica*. (Ministerio da Saude, Brasilia, Brasil, ed. 6, 2005), pp. 816. Available at: [http://www.prosaude.org/publicacoes/guia/Guia\\_Vig\\_Epid\\_novo2.pdf](http://www.prosaude.org/publicacoes/guia/Guia_Vig_Epid_novo2.pdf)
81. L. M. A. Camargo, M. U. Ferreira, H. Krieger, E. P. Camargo, L. P. Silva, Unstable hypoendemic malaria in Rondonia (Western Amazon Region, Brazil): epidemic outbreaks and work-associated incidence in an agro-industrial rural settlement. *Am. J. Trop. Med. Hyg.* **51**, (1994).
82. P. Barreto, C. Souza Jr., R. Nogueron, A. Anderson, R. Salomao, *Human Pressure on the Brazilian Amazon Forests*. G. Mock, Ed., (World Resources Institute, Belem, 2006). Available at: [http://www.globalforestwatch.org/common/pdf/Human\\_Pressure\\_Final\\_English.pdf](http://www.globalforestwatch.org/common/pdf/Human_Pressure_Final_English.pdf)
83. IBGE, *IBGE: Instituto Brasileiro de Geografia e Estatistica*. (2010). Available at: <http://www.ibge.gov.br>. Accessed 2012 March 19
84. INPE, *Projeto Prodes: monitoramento da Floresta Amazonica Brasileira por Satelite*. (Sao Jose dos Campos, SP, Brasil, 2010). Available at: [www.obt.inpe.br/prodes/](http://www.obt.inpe.br/prodes/). Accessed 2012 March 19
85. NASA, *3B43: Monthly 0.25 x 0.25 degree merged TRMM and other sources estimates*. (Maryland, USA., 2010). Available at: <http://mirador.gsfc.nasa.gov/cgi-bin/mirador/presentNavigation.pl?tree=project&project=TRMM&dataGroup=Gripped&dataset=3B43:%20Monthly%200.25%20x%200.25%20degree%20merged%20TRMM%20and%20other%20sources%20estimates&version=006>

86. L. E. O. C. Aragao *et al.*, Spatial patterns and fire response of recent Amazonian droughts. *Geophys Res Lett* **34**, (2007).
87. O. L. Phillips *et al.*, Drought sensitivity of the Amazon rainforest. *Science* **323**, 1344 (2009).
88. S. L. Lewis, P. M. Brando, O. L. Phillips, G. M. F. van der Heijden, D. Nepstad, The 2010 Amazon drought. *Science* **331**, 554 (2011).
89. A. Gelman, Prior distributions for variance parameters in hierarchical models. *Bayesian Anal* **1**, 515 (2006).
90. J. S. Clark, *Models for Ecological Data*. (Princeton University Press, Princeton, 2007). Available
91. A. Gelman, D. B. Rubin, Inference from iterative simulation using multiple sequences. *Stat Sci* **7**, 457 (1992).
92. B. S. Soares-Filho *et al.*, Modelling conservation in the Amazon basin. *Nature* **440**, 520 (2006).
93. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, *Bayesian Data Analysis*. (Chapman & Hall, London, 2003). Available
94. F. S. M. Barros, N. A. Honorio, M. E. Arruda, Survivorship of *Anopheles darlingi* (Diptera: Culicidae) in relation with malaria incidence in the Brazilian Amazon. *PLOS One* **6**, (2011).
95. F. S. M. Barros, N. A. Honorio, M. E. Arruda, Temporal and spatial distribution of malaria within an agricultural settlement of the Brazilian Amazon. *J Vector Ecol* **36**, 159 (2011).
96. C. A. Peres, Effects of subsistence hunting on vertebrate community structure in Amazonian Forests. *Conserv Biol* **14**, 240 (2000).

97. C. A. Peres, P. M. Dolman, Density compensation in neotropical primate communities: evidence from 56 hunted and nonhunted Amazonian forests of varying productivity. *Oecologia* **122**, 175 (2000).
98. K. H. Redford, The empty forest. *BioScience* **42**, 412 (1992).
99. L. M. Deane, Malaria vectors in Brazil. *Mem. Inst. Oswaldo Cruz* **81**, 5 (1986).
100. M. M. Povaia *et al.*, Malaria vectors, epidemiology, and the re-emergence of *Anopheles darlingi* in Belem, Para, Brazil. *J Med Entomol* **40**, 379 (2003).
101. D. Nepstad *et al.*, The end of deforestation in the Brazilian Amazon. *Science* **326**, 1350 (2009).
102. P. Tiwari, A. N. Sharma, An assessment of socio-demographic correlates associated with cases of maternal malaria. *J Hum Ecol* **12**, 371 (2001).
103. A. Unger, L. W. Riley, Slum health: from understanding to action. *PLoS Med* **4**, (2007).
104. B. Soares-Filho *et al.*, Role of Brazilian Amazon protected areas in climate change mitigation. *Proc Natl Acad Sci* **107**, 10821 (2010).
105. L. N. Joppa, S. R. Larie, S. L. Pimm, On the protection of "protected areas". *Proc Natl Acad Sci* **105**, 6673 (2008).
106. P. J. Ferraro, M. M. Hanauer, K. R. E. Sims, Conditions associated with protected area success in conservation and poverty reduction. *Proc Natl Acad Sci* **108**, 13913 (2011).
107. L. N. Joppa, A. Pfaff, Global protected area impacts. *Proc. R. Soc. B.* **278**, 1633 (2011).

108. D. Nepstad *et al.*, Inhibition of Amazon deforestation and fire by parks and indigenous lands. *Conserv Biol* **20**, 65 (2006).
109. P. J. Ferraro, M. M. Hanauer, Protecting ecosystems and alleviating poverty with parks and reserves: 'win-win' or tradeoffs? *Environ. Resource Econ.* **48**, 269 (2011).
110. G. Kindermann *et al.*, Global cost estimates of reducing carbon emissions through avoided deforestation. *Proc Natl Acad Sci* **105**, 10302 (2008).
111. T. Bousema *et al.*, Hitting hotspots: spatial targeting of malaria for control and elimination. *PLoS Med* **9**, (2012).
112. A. O. Brandao Jr., C. M. S. Souza Jr., Mapping unofficial roads with Landsat images: a new tool to improve the monitoring of the Brazilian Amazon rainforest. *Int J Remote Sens* **27**, 177 (2006).
113. D. Kaimowitz, B. Mertens, S. Wunder, P. Pacheco, *Hamburger connection fuels Amazon destruction*. (Center for International Forest Research, Bogor, Indonesia, 2004). Available at: [www.cifor.org/publications/pdf\\_files/media/Amazon.pdf](http://www.cifor.org/publications/pdf_files/media/Amazon.pdf)
114. P. M. Fearnside, Deforestation in Brazilian Amazonia: history, rates, and consequences. *Conserv Biol* **19**, 680 (2005).
115. CBD, PAHO/WHO, *Background Paper for the Regional Workshop on the Inter-Linkages between Human Health and Biodiversity in the Americas*. (Convention on Biological Diversity and Pan American Health Organization, 2012). Available at: <http://www.cbd.int/doc/meetings/health/wshb-am-01/other/wshb-am-01-interlink-en.pdf>
116. M. A. Barrett, T. A. Bouley, A. H. Stoertz, R. W. Stoertz, Integrating a One Health approach in education to address global health and sustainability challenges. *Frontiers in Ecology and Environment* **9**, 239 (2011).

117. EcoHealth Alliance. (EcoHealth Alliance), 2012. Available at: [http://www.ecohealthalliance.org/programs/26-consortium\\_for\\_conservation\\_medicine](http://www.ecohealthalliance.org/programs/26-consortium_for_conservation_medicine). Accessed 2012 Dec 13
118. A. O. Mala *et al.*, Dry season ecology of *Anopheles gambiae* complex mosquitoes at larval habitats in two traditionally semi-arid villages in Baringo, Kenya. *Parasites and Vectors* **4**, (2011).
119. J. M. Mwangangi *et al.*, *Anopheles* larval abundance and diversity in three rice agro-village complexes Mwea irrigation scheme, central Kenya. *Malar J* **9**, (2010).
120. E. J. Muturi *et al.*, Larval habitat dynamics and diversity of *Culex* mosquitoes in rice agro-ecosystem in Mwea, Kenya. *Am. J. Trop. Med. Hyg.* **76**, 95 (2007).
121. B. A. Ndenga, J. A. Simbauni, J. P. Mbugi, A. K. Githeko, Physical, chemical and biological characteristics in habitats of high and low presence of Anopheline larvae in Western Kenya Highlands. *PLOS One* **7**, (2012).
122. O. Kenea, M. Balkew, T. Gebre-Michael, Environmental factors associated with larval habitats of anopheline mosquitoes (Diptera: Culicidae) in irrigation and major drainage areas in the middle course of the Rift Valley, central Ethiopia. *Journal of Vector Borne Disease* **48**, 85 (2011).
123. Y. A. Afrane, B. W. Lawson, R. Brenya, T. Kruppa, G. Yan, The ecology of mosquitoes in an irrigated vegetable farm in Kumasi, Ghana: abundance, productivity and survivorship. *Parasites and Vectors* **5**, (2012).
124. M. A. Sattler *et al.*, Habitat characterization and spatial distribution of *Anopheles* sp. mosquito larvae in Dar es Salaam (Tanzania) during an extended dry period. *Malar J* **4**, (2005).
125. B. Matthys *et al.*, Urban agricultural land use and characterization of mosquito larval habitats in a medium-sized town of Cote d'Ivoire. *J Vector Ecol* **31**, 319 (2006).

126. X.-B. Liu *et al.*, Random repeated cross sectional study on breeding site characterization of *Anopheles sinensis* larvae in distinct villages of Yongcheng City, People's Republic of China. *Parasites and Vectors* **5**, (2012).
127. S. O. Vanwambeke *et al.*, Landscape and land cover factors influence the presence of *Aedes* and *Anopheles* larvae. *J Med Entomol* **44**, 133 (2007).
128. D. Vezzani, A. Rubio, S. M. Velazquez, N. Schweigmann, T. Wiegand, Detailed assessment of microhabitat suitability for *Aedes aegypti* (Diptera: Culicidae) in Buenos Aires, Argentina. *Acta Trop* **95**, 123 (2005).
129. C. H. Davis, D. P. MacKinnon, A. Schultz, I. Sandler, Cumulative risk and population attributable fraction in prevention. *Journal of Clinical Child and Adolescent Psychology* **32**, 228 (2003).
130. C. J. M. Koenraadt *et al.*, Spatial and temporal patterns in pupal and adult production of the dengue vector *Aedes aegypti* in Kamphaeng Phet, Thailand. *Am. J. Trop. Med. Hyg.* **79**, 230 (2008).
131. A. C. Morrison *et al.*, Temporal and geographic patterns of *Aedes aegypti* (Diptera: Culicidae) production in Iquitos, Peru. *J Med Entomol* **41**, 1123 (2004).
132. D. Valle, J. Clark, Conservation efforts may increase malaria burden in the Brazilian Amazon. *PLOS One* **8**, e57519 (2013).
133. J. A. Guarda, C. R. Asayag, R. Witzig, Malaria reemergence in the Peruvian Amazon region. *Emerg Infect Dis* **5**, 209 (1999).
134. J. Oliveira-Ferreira, R. Lourenco-de-Oliveira, A. Teva, L. M. Deane, C. T. Daniel-Ribeiro, Natural malaria infections in anophelines in Rondonia state Brazilian Amazon. *Am. J. Trop. Med. Hyg.* **43**, (1990).
135. C. Flores-Mendoza, R. Fernandez, K. S. Escobedo-Vargas, Q. Vela-Perez, G. B. Schoeler, Natural Plasmodium infections in *Anopheles darlingi* and *Anopheles benarrochi* (Diptera: Culicidae) from Eastern Peru. *J Med Entomol* **41**, 489 (2004).

136. J. Hayes, G. Calderon, R. Falcon, V. Zambrano, Newly incriminated anopheline vectors of human malaria parasites in Junin Department, Peru. *J Am Mosq Control Assoc.* **3**, 418 (1987).
137. J. N. Ijumba, S. W. Lindsay, Impact of irrigation on malaria in Africa: paddies paradox. *Medical and Veterinary Entomology* **15**, 1 (2001).
138. S. Maki, R. Kalliola, K. Vuorinen, Road construction in the Peruvian Amazon: process, causes and consequences. *Environmental Conservation* **28**, 199 (2001).
139. A. Y. Vittor, Johns Hopkins University (2003).
140. D. G. T. Denison, C. C. Holmes, B. K. Mallick, A. F. M. Smith, *Bayesian methods for nonlinear classification and regression*. (John Wiley & Sons, England, 2002). Available
141. U. Haque *et al.*, Malaria prevalence in endemic districts of Bangladesh. *PLOS One* **4**, (2009).
142. M. H. Craig, B. L. Sharp, M. L. H. Mabaso, I. Kleinschmidt, Developing a spatial-statistical model and map of historical malaria prevalence in Botswana using a staged variable selection procedure. *International Journal of Health Geographics* **6**, (2007).
143. I. R. F. Elyazar *et al.*, Plasmodium vivax malaria endemicity in Indonesia in 2010. *PLOS One* **7**, (2012).
144. P. W. Gething *et al.*, A new world malaria map: Plasmodium falciparum endemicity in 2010. *Malar J* **10**, (2011).
145. L. Gosoni, P. Vounatsou, N. Sogoba, N. Maire, T. Smith, Mapping malaria risk in West Africa using a Bayesian nonparametric non-stationary model. *Computational Statistics & Data Analysis* **53**, 3358 (2009).

146. P. Diggle, R. Moyeed, B. Rowlingson, M. Thomson, Childhood malaria in the Gambia: a case-study in model-based geostatistics. *Appl. Statist.* **51**, 493 (2002).
147. H. L. Reid, U. Haque, S. Roy, N. Islam, A. C. A. Clements, Characterizing the spatial and temporal variation of malaria incidence in Bangladesh, 2007. *Malar J* **11**, (2012).
148. N. Silva *et al.*, Epidemiology and control of frontier malaria in Brazil: lessons from community-based studies in rural Amazonia. *Trans R Soc Trop Med Hyg* **104**, 343 (2010).
149. A. Taye, M. Hadis, N. Adugna, D. Tilahun, R. A. Wirtz, Biting behavior and Plasmodium infection rates of Anopheles arabiensis from Sille, Ethiopia. *Acta Trop* **97**, 50 (2006).
150. S. W. Lindsay, L. Parson, C. J. Thomas, Mapping the range and relative abundance of the two principal African malaria vectors, Anopheles gambiae sensu stricto and An. arabiensis, using climate data. *Proc. R. Soc. Lond. B* **265**, 847 (1998).
151. R. S. Levine, A. T. Peterson, M. Q. Benedict, Geographic and ecologic distributions of the Anopheles gambiae complex predicted using a genetic algorithm. *Am. J. Trop. Med. Hyg.* **70**, 105 (2004).
152. J. Yasuoka, R. Levins, Impact of deforestation and agricultural development on anopheline ecology and malaria epidemiology. *Am. J. Trop. Med. Hyg.* **76**, 450 (2007).
153. T. F. Silva-do-Nascimento, R. C. Wilkerson, R. Lourenco-de-Oliveira, F. A. Monteiro, Molecular confirmation of the specific status of Anopheles halophylus (Diptera: Culicidae) and evidence of a new cryptic species within An. triannulatus in Central Brazil. *J Med Entomol* **43**, 455 (2006).

154. T. F. Silva-do-Nascimento, R. Lourenco-de-Oliveira, Diverse population dynamics of three Anopheles species belonging to the Triannulatus Complex (Diptera: Culicidae). *Mem. Inst. Oswaldo Cruz* **102**, 975 (2007).
155. T. A. Klein, J. B. P. Lima, M. S. Tada, Comparative susceptibility of anopheline mosquitoes to Plasmodium falciparum in Rondonia, Brazil. *Am. J. Trop. Med. Hyg.* **44**, 598 (1991).
156. J. D. Charlwood, Biological variation in Anopheles darlingi Root. *Mem. Inst. Oswaldo Cruz* **91**, 391 (1996).
157. R. Aguas, L. J. White, R. W. Snow, M. G. M. Gomes, Prospects for malaria eradication in Sub-Saharan Africa. *PLoS ONE* **3**, e1767 (2008).
158. C. E. A. Coimbra, Human factors in the epidemiology of malaria in the Brazilian Amazon. *Human Organization* **47**, 254 (1988).
159. L. E. O. C. Aragao *et al.*, Spatial patterns and fire response of recent Amazonian droughts. *Geophysical Research Letters* **34**, (2007).
160. J. C. Derber, D. F. Parrish, S. J. Lord, The New Global Operational Analysis System at the National-Meteorological-Center. *Weather and Forecasting* **6**, 538 (1991).
161. G. E. Liston, K. Elder, A Meteorological Distribution System for High-Resolution Terrestrial Modeling (MicroMet). *Journal of Hydrometeorology* **7**, 217 (2006).
162. G. J. Huffman *et al.*, The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *Journal of Hydrometeorology* **8**, 38 (2007).
163. F. Chen *et al.*, Modeling of land surface evaporation by four schemes and comparison with FIFE observations. *Journal of Geophysical Research-Atmospheres* **101**, 7251 (1996).

164. M. B. Ek *et al.*, Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *Journal of Geophysical Research-Atmospheres* **108**, (2003).
165. S. V. Kumar *et al.*, Land information system: An interoperable framework for high resolution land surface modeling. *Environmental Modelling & Software* **21**, 1402 (2006).

## **Biography**

Denis Valle was born in December 10<sup>th</sup> (1979) in Sao Paulo (Brasil). He studied Forestry in the University of Sao Paulo from 1998-2003. After finishing his studies there, he worked for three years as a research assistant at a Brazilian NGO called IMAZON. In 2006, he started a Master of Science in the department of Forest Resources and Conservation at the University of Florida under the supervision of Dr. Christina Staudhammer, obtaining a minor in Statistics. In 2008, Denis Valle enrolled in the PhD program in Ecology at Duke University under the supervision of Dr. James Clark. While at Duke, Denis Valle also pursued a concurrent Master of Science in statistics. Denis Valle was recipient of the 2012-2013 Katherine Goodman Stern Fellowship at Duke.