Grant agreement no. 243964

QWeCI

Quantifying Weather and Climate Impacts on Health in Developing Countries

Deliverable 1.1.b Report on current climate sensitivities of diseases and projections of future distributions including a mapped output on the project website

Start date of project: 1st February 2010 Duration: 42 months

Lead contractor: UNILIV
Coordinator of deliverable: UNILIV
Evolution of deliverable

Due date : M38
Date of first draft : 8 July 2013
Start of review : 15 July 2013
Deliverable accepted : 30 July 2013

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination Level		
PU	Public	PU
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Introduction to the EID2

A database of all human and animal pathogens (the ENHanCEd Infectious Disease Database, EID2) has been completed as a part of the ENHanCE (www.liv.ac.uk/enhance) and QWeCI projects. The EID2 has been built using automated methodologies which extract information from the literature and from nucleotide sequences submitted to Genbank and other databases (accessible via the NCBI Taxonomy Database (http://www.ncbi.nlm.nih.gov/taxonomy) (for further details see . Pathogen entries within the EID2 are labelled with information on their source (i.e. host), where they are found (at the country-level) and when they were isolated. This information is linked to map systems and Google Earth (Figure 1), and enables us to study pathogens that are present in the target region for the QWeCI project, Africa.

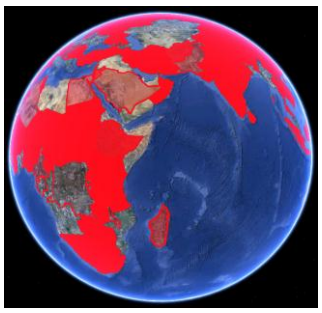


Figure 1. Example output from EID2 - the distribution of rabies virus according to the NCBI nucelotide database. Red colour indicates sequences of rabies virus in the NCBI nucleotide database from the appropriate country; red border, 1-9 sequences; red fill, 10 or more sequences.

The capability to spatially map disease or pathogen data has been built into the EID2 database at the same spatial resolution as fields of climate data and some host population data. Both the presence-only and climate/population data can be exported for each pathogen from within the EID2.

The information on the presence of pathogens provided by the EID2, the climate and host population data were used to perform predictive presence/absence disease modelling for some of the diseases (or major pathogens causing diseases) studied within the QWeCI project. Modelling exercises were undertaken for *Plasmodium falciparum*, Rift Valley Fever, and the tick-borne pathogen *Babesia bigemina*. Two different types of modelling exercises were developed at different spatial scales: (1) at the country-level and (2) at a 0.25 degree square resolution. These exercises are presented as two separate studies within this deliverable report.

References

McIntyre, K.M., Setzkorn, C., Wardeh, M., Hepworth, P.J., Radford, A.D., Baylis, M., 2013. What, where and weather? Integrating open-source taxonomic, spatial and climatological information into a comprehensive database of livestock infections. In, Proceedings of the Annual Meeting 2013, Society for Veterinary Epidemiology and Preventative Medicine, Madrid, Spain.

Study 1. Presence/absence disease modelling at the country-level using population density and bioclimatic data

Abstract
The world occurrence of three diseases: Rift Valley Fever, *Babesia bigemina* and *Plasmodium falciparum,* was modelled at the country-level using two presence/absence algorithms: generalized linear models (GLM) and multivariate adaptive regression splines (MARS). Models were evaluated using a k-fold cross-validation procedure (k=5) in terms of their predictive skill as given by the ROC skill area metric (RSA). The independent contribution of the different variables was also assessed in the context of GLM modelling using a technique involving hierarchical partitioning in order to gain an insight into the importance of the climatic/human/animal factors in explaining disease occurrence. Our results indicate that no added skill was attained with the use of the more sophisticated MARS technique. Model skill was poor in the case of *B. bigemina* infection. Models for *P. falciparum* and particularly Rift Valley Fever attained moderate/good skill, indicating the potential usefulness of the models developed. Regarding variable importance, the results indicate that at the country-level, diseases can be modelled using bioclimatic variables as predictors, with little or no added benefit from the inclusion of host density predictors.The exploitation of the data stored in EID2 has enabled a straightforward development of the models presented, leaving the door opened to further advances in disease distribution modelling using this database.

Methods

Modelling algorithms

We have tested two different algorithms for model development: generalized linear models (GLM, McCullagh and Nelder, 1989) and multivariate adaptive regression splines (MARS, Friedman, 1991). GLMs constitute a commonly used parametric method, thus constituting an adequate tool for benchmarking. MARS, in contrast, is a non–parametric method for regression which approximates the underlying function through a set of adaptive piecewise linear regressions, known as basis functions. Unlike GLMs, MARS is able to model nonlinearities in the data. A comparative study of these algorithms in the context of species' distribution modelling is described in Bedia et al. (2011) and Bedia et al. (2013).
Presence data for pathogens or diseases was provided from within the EID2 at the country-level for all countries.

Presence data for pathogens

Presence data for pathogens or diseases was provided from within the EID2 at the country-level for all countries. For more information see study 2 (0.25 degree square resolution modelling) and McIntyre et al. (2013).

Explanatory variables

In this work we considered a set of 14 bioclimatic variables commonly used in ecological modelling (see Table 1). The main advantage of these bioclimatic variables over other seasonal indices is that they are calculated and applied independently of the country's hemisphere. In this study, we used only bioclimatic variables which could be calculated using mean temperature and precipitation data, as minimum and maximum temperatures, which allow for the calculation of other indices, were not available. In order to avoid redundancy, we eliminated from the analysis bioclimatic variables which yielded correlation values above 0.90 (Spearman's rho coefficient) in the pairwise cross-correlation matrix (Figure 1). A threshold of 0.90 is conservative, and was chosen in order to keep other variables that, although also highly correlated, may still provide some useful additional information. The final dataset of bioclimatic variables is presented in Table 1.

Table 1. Summary of explanatory variables. Bioclimatic variables marked with an asterisk were eliminated after correlation analysis. The super-indices in the population variables indicate: 1 = used within the *Plasmodium falciparum* model, 2 = used within the Rift Valley Fever model, 3 = used within the *Babesia bigemina* model.

Code	Variable definition
Bioclimatic variables	
bio1	Mean annual temp.
bio4	Temp. seasonality
bio8	Mean temp. of wettest quarter
bio9*	Mean temp. of driest quarter
bio10*	Mean temp. of warmest quarter

bio11*	Mean temp. of coldest quarter
bio12	Annual precip.
bio13*	Precip. of wettest month
bio14	Precip. of driest month
bio15	Seasonality of precip.
bio16*	Precip. of wettest quarter
bio17*	Precip. of driest quarter
bio18	Precip. of warmest quarter
bio19	Precip. of coldest quarter
Population variables	
Human pop[1;2]	Human population density
Sheep pop[2]	Sheep population density
Goat pop[2]	Goat population density
Catbuf pop[2;3]	Cattle/Buffalo population density

All data were interpolated to a common regular grid of 0.25 degree square spatial resolution. The climate data was based on the CRUTS2.1 gridded data based on 1950-2000 climatology[1].

Host population density data

Host population density data (see Table 1) was included for the main hosts of each of the three diseases/pathogens when it was available. The animal density estimates had an original resolution of 0.05 degree, corresponding to predicted outputs for 2005[2]. The human population density was obtained from the gridded population of the world data set (GPWv3[3]) for 2005, with an original resolution of 0.25 degrees.

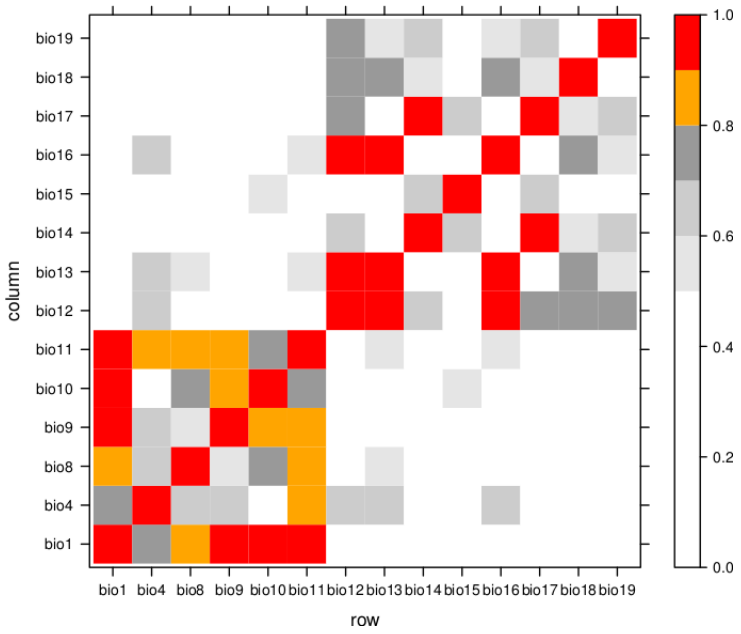


Figure 1. Correlation matrix of the bioclimatic variables considered in this study. Variables with coefficients above 0.9 have been considered redundant (see Table 1 for details).

Finally, all data were averaged at the country-level (N = 176 world countries) and standardized, prior to model building.

Model assessment

We performed a k-fold cross-validation of the models, with k=5 stratified randomly split subsets of presence/absence, each of them containing an approximately equal number of presences and absences. Model skill was assessed by computing the ROC curves for each model and calculating the corresponding RSA. We also tested the leave-one out

1 http://www.cru.uea.ac.uk/~timm/grid/CRU_TS_2_1.html
2 http://www.fao.org/Ag/AGAInfo/resources/en/glw/GLW_dens.html
3 http://sedac.ciesin.columbia.edu/data/set/gpw-v3-population-density-future-estimates

cross-validation procedure, attaining similar RSA values to those from the k-fold method. Thus, we selected the latter procedure, because it makes possible to provide a measure of RSA spread.

Variable importance assessment

The rationale behind the use of GLMs and MARS lies in the flexible and robust framework used to estimate variable importance, which allows disentanglement of the roles of bioclimatic and human/animal population factors in explaining disease distributions. Although non-linear techniques such as MARS may lead to models of improved predictive accuracy (Elith et al., 2006; Bedia et al., 2011), they may also eventually obscure the actual contribution of each variable due to their greater complexity. In contrast, GLMs provide a flexible and robust framework for assessing the statistical significance of explanatory variables and estimation of their importance, and a straightforward model interpretation at a low computational cost.

In order to estimate variable importance in the context of logistic regression modelling, we have applied a method of hierarchical partitioning, by which the independent effect of each variable is calculated by comparing the fit of all models containing a particular variable to the fit of all nested models lacking that variable (Chevan and Sutherland, 1991). For instance, for variable X_1, its importance, I, would be calculated as follows:

$$I_{x1} = \sum_{i=0}^{k-1} \frac{\sum (r^2_{y,X_1 X_h} - r^2_{y,X_h}) / \binom{k-1}{i}}{k}$$

where X_h is any subset of i predictors from which X_1 is excluded. As a result, the variance shared by two or more correlated predictors can be partitioned into the variance attributed to each predictor. This method provides a robust assessment of variable importance and has been shown to outperform other methods used for variable importance estimation in the context of regression analysis, after the removal of spurious variables (Murray and Conner, 2009).

All the analyses were conducted in the R language and environment for statistical computing (R Development Core Team, 2012). The hierarchical partitioning work was undertaken using the R package hier.part (Walsh and Mac Nally, 2013). For the MARS models, we used the implementation of the algorithm included in the R package earth (Milborrow, 2013).

Results

Both modelling methods (GLM and MARS) yielded similar results in terms of RSA (Table 2). Thus, we selected generalized linear models (GLMs) as the preferred technique to use. The inclusion of the population-related variables in the dataset of predictors (denoted as popclim in Table 2) provided marginal or null increments of RSA in the models, revealing the dominance of climatic factors in modelling disease occurrence at the country-level. The predicted probability world maps at the country level of the climatic GLM models are presented in Figure 2, considering the two diseases attaining an acceptable model performance (i.e. Rift Valley Fever and *P. falciparum*).

Table 2. ROC skill area (RSA) of the 5–fold cross–validation models (mean sigma).

	GLM		**MARS**	
	popclim	*clim*	*popclim*	*clim*
Babesia bigemina	0.60_0.14	0.57_0.13	0.61_0.22	0.65_0.16
Plasmodium falciparum	0.76_0.03	0.76_0.04	0.76_0.09	0.77_0.06
Rift Valley Fever	0.83_0.03	0.87_0.03	0.83_0.11	0.69_0.09

The importance of bioclimatic variables was confirmed by the relative importance of some bioclimatic variables compared to population-related ones (Figure 3). In the case of *B. bigemina*, the most important variables were bio14 (precipitation of driest month) and bio4 (temperature seasonality). Although the cattle/buffalo population density variable was relatively important in terms of its mean, the spread among the 5-fold dataset was very large, and therefore its effect must be considered marginal. In addition, the model of *B. bigemina* attained a very low RSA, and therefore none of the variables were critical in explaining its occurrence at the country-level.

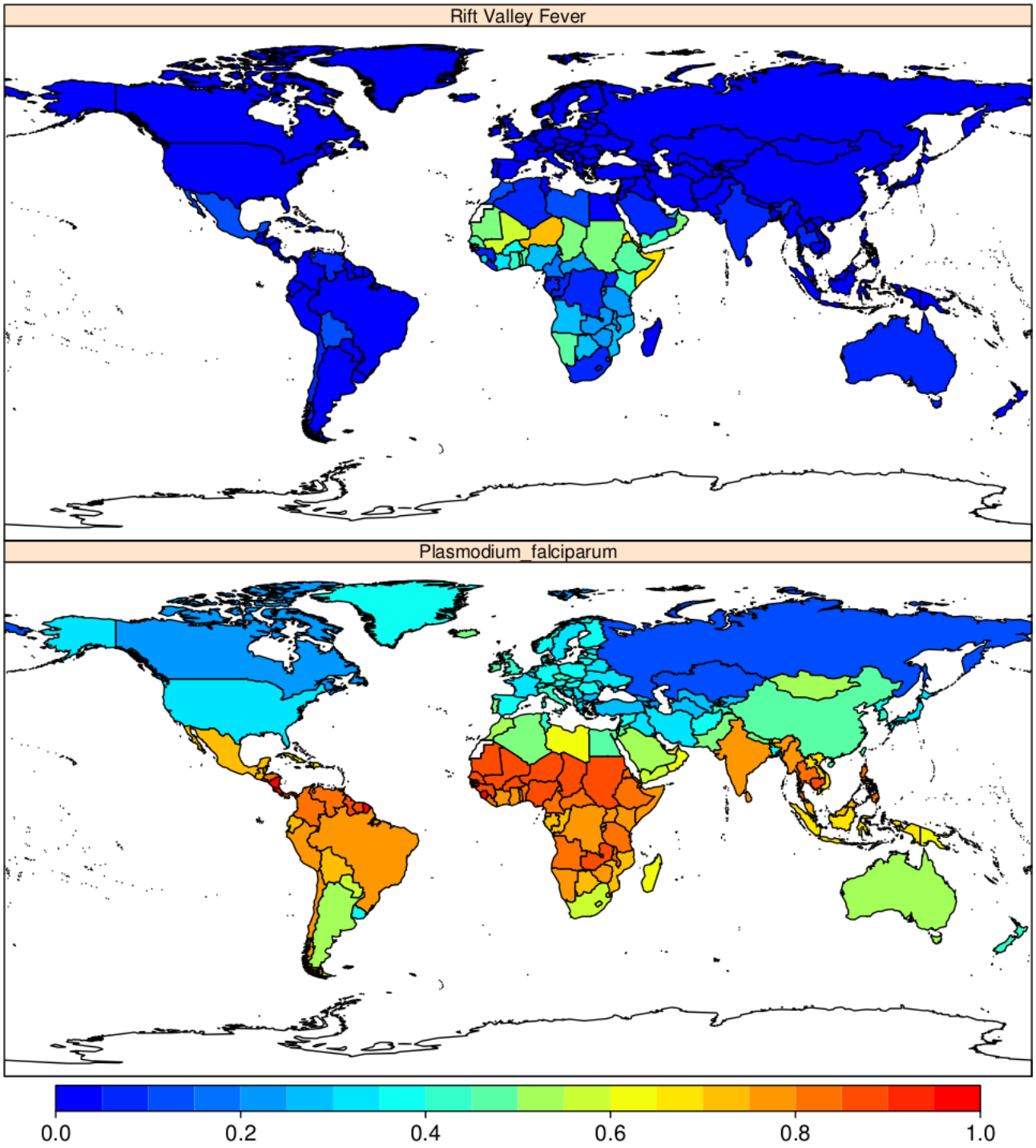


Figure 2: Predicted probability maps of the GLM occurrence models for Rift Valley Fever (RSA = 0.83) and *Plasmodium falciparum* (RSA = 0.76) at the world country level. Results correspond to the climate-only models. Note that the colour legend has been designed in order to aid in the comparison with the spatial distribution of *Plasmodium falciparum* malaria endemicity in 2010 presented in Gething et al. (2011) (see also Figure 4).

In the case of *P. falciparum*, the most important variable was bio4 (temperature seasonality), with sizeable contributions of bio1 (mean annual temperature), bio15 (seasonality of precipitation) and bio8 (mean temperature of wettest quarter), the latter with a larger multi-model spread. Again, human population density made a marginal contribution to total explained variance, and did not improve model skill.

The most important variables in the Rift Valley Fever model were bio14 (precipitation of driest month) and bio15 (seasonality of precipitation), although with a larger spread than bio1 and bio8; these variables were also relatively important. The human and animal population variables all made a small contribution to total explained variance.

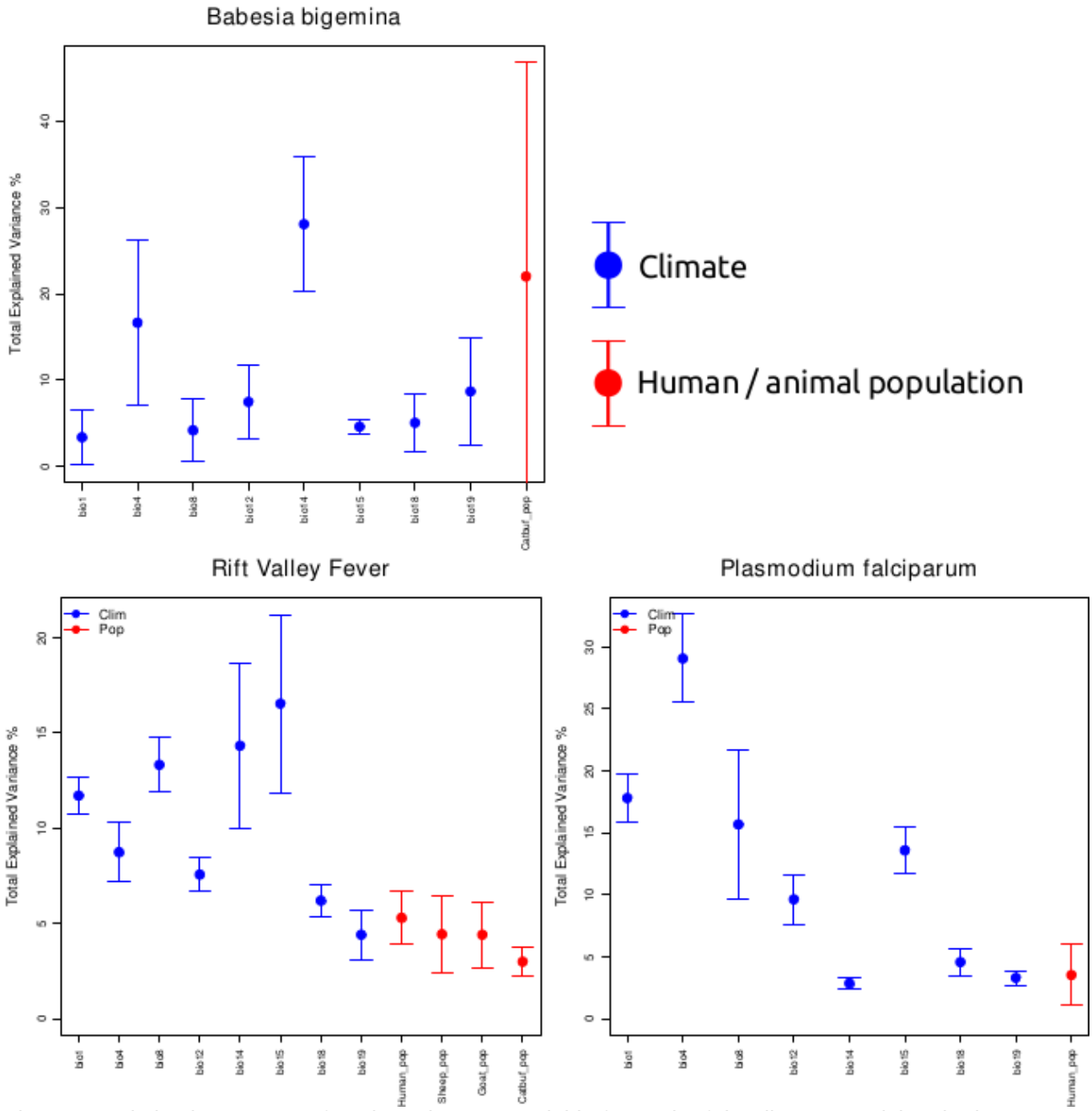


Figure 3. Relative importance of each explanatory variable for each of the disease models. The importance is expressed as a percentage of the total explained variance, considering the independent effect of each variable according to the hierarchical partitioning approach (see Section 1.4). Points correspond to the mean of the 5-fold models and the vertical bars to the standard deviation. See Table 1 for variable code definition.

The comparison of our results of *P. falciparum* with the endemicity map for the year 2010 published by Gething et al. (2011) reveals a good agreement between the areas where the disease is present and the predicted probabilities of occurrence at the country level. In fact, after data aggregation by countries (N=77), only 6 countries (7.8%) where the disease is reported are given low suitability values by the model (i.e., below the probability threshold for positive case classification), and these correspond to low endemicity values below 10% (Figure 4).

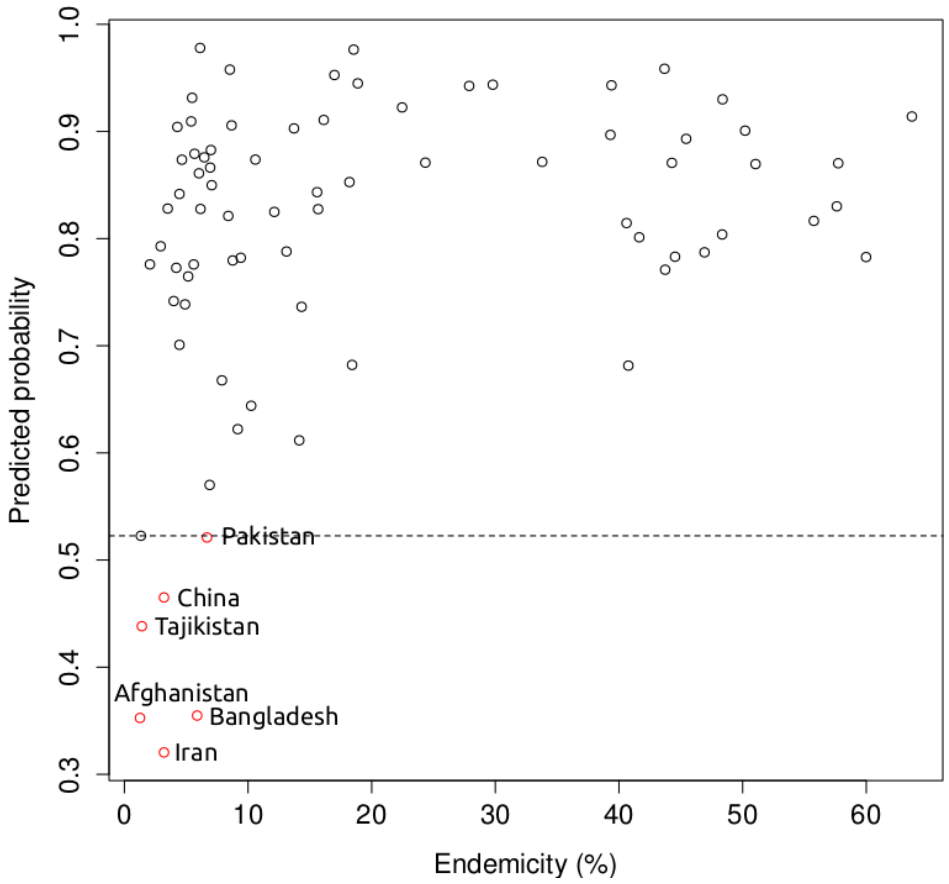


Figure 4: Endemicity levels averaged by countries where *Plasmodium falciparum* has been reported (Gething et al., 2011), versus predicted probabilities for the same countries (N = 77) according to the bioclimatic GLM model. The probability threshold for case classification is indicated by the horizontal line. The wrongly classified countries are indicated by the labelled red points.

References

BEDIA, J. AND BUSQU´E, J. AND GUTI´E RREZ, J. M., 2011: Predicting plant species distribution across an alpine rangeland in northern Spain: a comparison of probabilistic methods, Applied Vegetation Science, 14, 415–432.

BEDIA, J. AND HERRERA, S. AND GUTI´ERREZ, J. M., 2013: Dangers of using global bioclimatic datasets for ecological niche modeling. Limitations for future climate projections, Global and Planetary Change, 107, 1–12, doi:10.1016/j.gloplacha.2013.
04.005.

CHEVAN, A. AND SUTHERLAND, M., 1991: Hierarchical Partitioning, The American Statistician, 45, 90–96.

ELITH, J. AND GRAHAM, C. H. AND ANDERSON, R. P. AND DUDIK, M. AND FERRIER, S. AND GUISAN, A. AND HIJMANS, R. J. AND HUETTMANN, F. AND LEATHWICK, J. R. AND LEHMANN, A. AND LI, J. AND LOHMANN, L. G. AND LOISELLE, B. A. AND MANION, G. AND MORITZ, C. AND NAKAMURA, M. AND NAKAZAWA, Y. AND OVERTON, J. M. AND PETERSON, A. T. AND PHILLIPS, S. J. AND RICHARDSON, K. AND SCACHETTI-PEREIRA, R. AND SCHAPIRE, R. E. AND SOBERON, J. AND WILLIAMS, S. AND WISZ, M. S. AND ZIMMERMANN, N. E., 2006: Novel methods improve prediction of species' distributions from occurrence data, Ecography, 29, 129–151.

FRIEDMAN, J. H., 1991: Multivariate adaptive regression splines, Annals of Statistics, 19, 1–67.

MCINTYRE, K.M., SETZKORN, C., WARDEH, M., HEPWORTH, P.J., RADFORD, A.D. AND BAYLIS, M., 2013: What, where and weather? Integrating open-source taxonomic, spatial and climatological information into a comprehensive database of livestock infections. In, Proceedings of the Annual Meeting, Society for Veterinary Epidemiology and Preventative Medicine, Madrid, Spain, 143-159. ISBN: 978-0-948073-20-5.

MILBORROW, S., 2013. DERIVED FROM MDA:MARS BY TREVOR HASTIE AND ROB TIBSHIRANI. USES ALAN MILLER'S FORTRAN UTILITIES WITH THOMAS LUMLEY'S LEAPS WRAPPER.: earth: Multivariate Adaptive Regression Spline Models, http://CRAN.R-project.org/package=earth, r package version 3.2-6.

MCCULLAGH, P. AND NELDER, J., 1989: Generalized linear models, Chapman & Hall, London.
MURRAY, K. AND CONNER, M.M., 2009: Methods to quantify variable importance: implications for the analysis of noisy ecological data, Ecology, 90, 348–355.
R DEVELOPMENT CORE TEAM: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, http://www.R-project.org, ISBN 3-900051-07-0, 2012.
CHRIS WALSH AND RALPH MAC NALLY: hier.part: Hierarchical Partitioning, http://CRAN.R-project.org/package=hier.part, r package version 1.0-4, 2013.

Study 2. Presence/absence modelling at a 0.25 degree square resolution using an expectation-maximisation algorithm technique applied to generalised additive modelling.

Introduction

The presence or absence of pathogens was modelled at a 0.25 degree square resolution using an expectation-maximisation (EM) algorithm technique (see Ward et al., 2009) which was applied to generalised additive modelling (GAM) with a binomial logistic error distribution. Presence was modelled for three pathogens with different extents of presence across the African continent. *Plasmodium falciparum* was examined as a pathogen with a very widespread distribution, Rift Valley Fever has a spatially-limited distribution, and *Babesia bigemina* has a localised distribution.

Methods

Presence data for pathogens

Presence data for pathogens or diseases was provided from within the ENHanCEd Infectious Disease Database (EID2) at the country-level for all countries on the African continent.

Within the EID2, specific information on pathogens occurring within a country (an 'organism-country interaction') was mined from meta-data held within the NCBI Nucleotide database (National Center for Biotechnology Information, 2012b); such information was treated as a 'gold standard' within the database. The data-mining was undertaken by searching the meta-data for entries describing pathogen infection occurring in a specific country. The last update from the nucleotide database was undertaken in December 2011. A further source of information came from automated searches of the PubMed database (National Center for Biotechnology Information, 2012c) and the NCBI MeSH library (National Center for Biotechnology Information, 2012a); when the name of an organism and the (minor subject) MeSH term for a country co-occurred within a certain number of publications, an assumption was made about the occurrence of that organism within that country. For further information see (McIntyre et al., 2013).

Explanatory variables

The same set of 14 bioclimatic variables was used for this modelling as for the MARS and GLM country-level modelling exercise (please see Study 1 and Table 1). Only bioclimatic variables which could be calculated using mean temperature and precipitation data were used. All data were interpolated to a common regular grid of 0.25 degree square spatial resolution. The climate data was based on the CRUTS2.1 gridded data over the time-period 1950-2000[4].

Table 1. Summary of explanatory variables. Bioclimatic variables marked with an asterisk were eliminated after correlation analysis. The super-indices in the population variables indicate: 1 = used within the *Plasmodium falciparum* model, 2 = used within the Rift Valley Fever model, 3 = used within the *Babesia bigemina* model.

Code	Variable definition
Bioclimatic variables	
bio1	Mean annual temp.
bio4	Temp. seasonality
bio8	Mean temp. of wettest quarter
bio9*	Mean temp. of driest quarter
bio10*	Mean temp. of warmest quarter

4 http://www.cru.uea.ac.uk/~timm/grid/CRU_TS_2_1.html

bio11*	Mean temp. of coldest quarter
bio12	Annual precip.
bio13*	Precip. of wettest month
bio14	Precip. of driest month
bio15	Seasonality of precip.
bio16*	Precip. of wettest quarter
bio17*	Precip. of driest quarter
bio18	Precip. of warmest quarter
bio19	Precip. of coldest quarter
Population variables	
Human pop[1;2]	Human population density
Sheep pop[2]	Sheep population density
Goat pop[2]	Goat population density
Catbuf pop[2;3]	Cattle/Buffalo population density

Host population density data

Host population density data was included for the main hosts of each of the three diseases/pathogens when it was available (see Table 1). The animal density estimates had an original resolution of 0.05 degree, corresponding to predicted outputs for 2005[5]. The human population density was obtained from the gridded population of the world data set (GPWv3[6]) for 2005, with an original resolution of 0.25 degrees.

The EM algorithm in GAM modelling

Code to run the EM algorithm technique applied to generalised additive modelling (EM-GAM) with a logistic error distribution was written for use within the R statistical package. This code uses pathogen presence, climate and host population density data exported from the EID2 to predict the presence or absence of pathogens given the climate and host density explanatory variables, according to different prior probabilities (*Pi*) of the likelihood of a detected absence of a pathogen (when it has not been recorded in the EID2) actually being a presence.

Using this technique, it is difficult to decide what value of *Pi* to use within the modelling exercise. In the modelling undertaken, multiple runs of models were initially used to test different (a) numbers of iterations and (b) *Pi* values incorporated within the EM-GAM technique, in order to see if it was possible to maximize the log likelihood value of the model and therefore to ascertain (a) the minimum number of iterations needed to allow each model to converge properly, and also (b) to try and see if an ideal (in which the log-likelihood value was maximized) prior probability of a detected absence actually being a presence could be predicted without prior knowledge. The minimum *Pi* values tested for each pathogen studied included: *Pi* = 0.1, 0.4, 0.5, 0.7, 0.9, 0.95, 0.98 and 0.99.

Adjusting for the surveillance of pathogens within the EM-GAM technique

The results of these exercises suggested that the logistic regression models converged after a certain number of iterations, but also that the prior probability of a pathogen absence actually being detected as a presence could not be predicted without some kind of (Bayesian) prior knowledge of the distribution of the pathogen, as previously suggested by Ward et al. (2009). An improved approach was therefore tested in which location-specific priors (*Pi*) would be estimated using a combined assessment of the surveillance effort for a pathogen in different countries and of the general surveillance of all pathogens in one country relative to other countries (Equation 1), utilising information on pathogen surveillance from the EID2 database and Figure 1):

$$Pi = (N_1 / N_t) * (1 - D_1 / D_t) \qquad (1)$$

Where N_1 is the number of African countries in which a certain pathogen is present, N_t is the total number of African countries, D_1 is the number of pathogen species detected in a certain African country, and D_t is the total number of pathogen species found in all African countries.

(a) (b) (c)

5 http://www.fao.org/Ag/AGAInfo/resources/en/glw/GLW_dens.html

6 http://sedac.ciesin.columbia.edu/data/set/gpw-v3-population-density-future-estimates

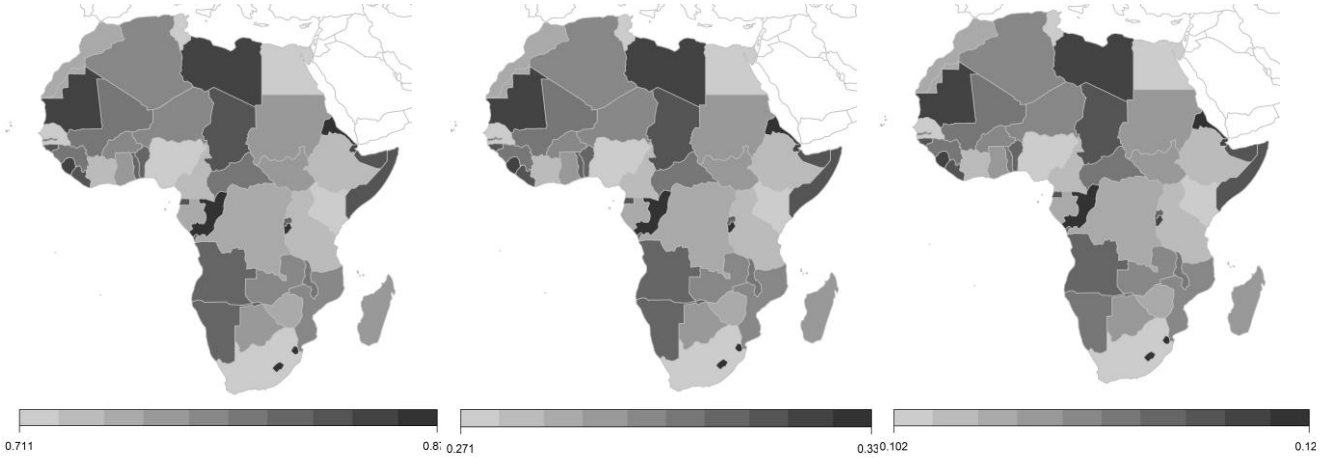


Figure 1. Surveillance effort for pathogens in African countries for (a) *Plasmodium falciparum*, (b) Rift Valley Fever, and (c) *Babesia bigemina*, calculated utilising information on pathogen surveillance provided by the EID2 database.

Multiple EM-GAM models were thereafter developed for *P. falciparum,* Rift Valley Fever and *B. bigemina* which incorporated the different bioclimatic and host density variables specified in Table 1, using different values of *Pi* (*Pi* = 0.0, 0.25, 0.50, 0.75, surveillance adjusted *Pi* value) to illustrate the impact of assumptions about pathogen presence given reported absence. The final model included all variables which significantly affected the variance of the model. Statistical significance was determined by a *P*-value of less than 0.05. Prior weights were assigned to the models for Rift Valley Fever and *B. bigemina*. A quasi-binomial family was used within the models for *P. falciparum* because the dispersion parameters were unknown. All models which used the surveillance adjusted *Pi* value utilised a quasi-binomial family. Likelihood estimates are not available for quasi-binomial models.

Results

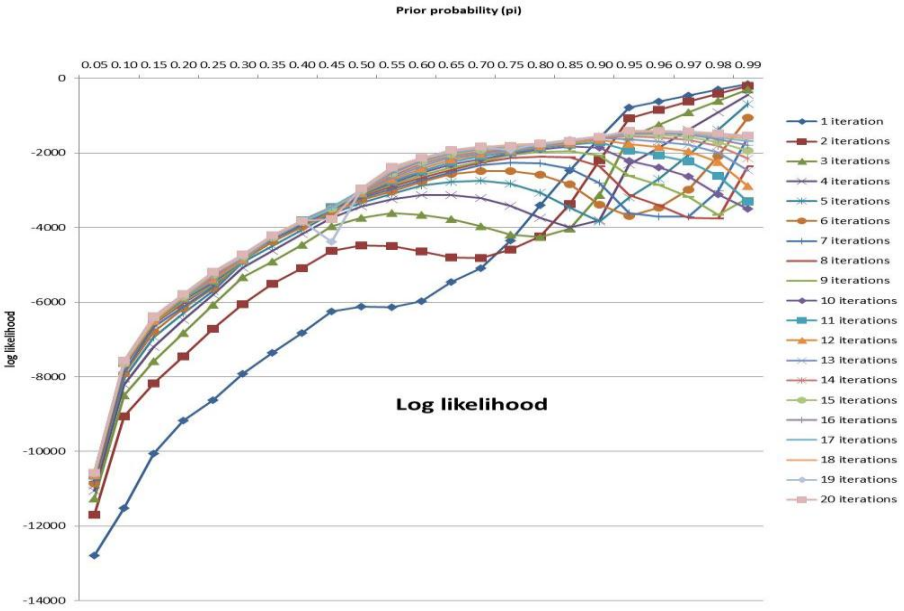
The use of different numbers of iterations within EM-GAM modelling

EM-GAM models including several simple climate terms were run for *P. falciparum* and Rift Valley Fever with different numbers of iterative procedures, from a single step model, to a model with 20 iterations. The model fit (measured by a higher log-likelihood value) converged as the number of iterations used within the model reached 20 (Figures 2a and b).

The use of different Pi values in EM-GAM modelling

When a series of *Pi* values were examined within the same EM-GAM models (including several simple climate terms), model fit increased as *Pi* approached 1 (where the pathogen is present in all regions) but did not peak at any point, suggesting that Pi can not be estimated using the data (Figures 2a and b), as previously suggested by Ward et al. (2009).

(a) *Plasmodium falciparum*



(b) Rift valley Fever

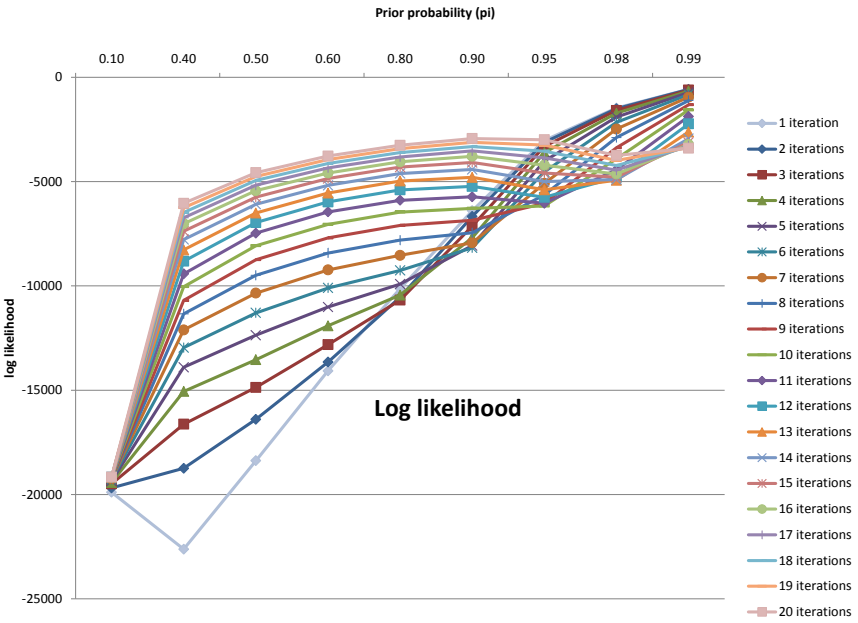


Figure 2. Log likelihood values for logistic regression models of (a) *Plasmodium falciparum* and (b) Rift Valley Fever which incorporated an expectation-maximisation algorithm technique applied to generalised additive modelling to predict presence or absence according to different prior probabilities (*Pi*) of the likelihood of a detected absence of a pathogen (when it has not been recorded in the EID2) actually being a presence. Values are depicted for models which examined different numbers of iterations and *Pi* values, to maximise the log likelihood value of the model. The model fit (measured by log-likelihood) converged as the number of iterations used within the model reached 20. In addition, model fit increased as *Pi* approached 1 (where the pathogen is present in all regions) but did not peak at any point, suggesting that *Pi* can not be estimated using the data, as previously suggested by Ward *et al.* (2009).

Final EM-GAM models run using different Pi *values*

All the bioclimatic variables (see Table 2) were statistically significant ($P<0.001$) and were included within the final EM-GAM models for *P. falciparum*, Rift Valley Fever and *B. bigemina* (Table 2 and Figure 3). The models for Rift Valley Fever and *B. bigemina* with the highest log-likelihood value used no iterative process (*Pi*=0.0) for locations where the pathogen was absent (Table 3). The predicted presence for models which used the surveillance adjusted *Pi* value were visually (in most cases), the ones in which the presence looked most similar to the *Pi*=0.0 models (Table 3 and Figure 3).

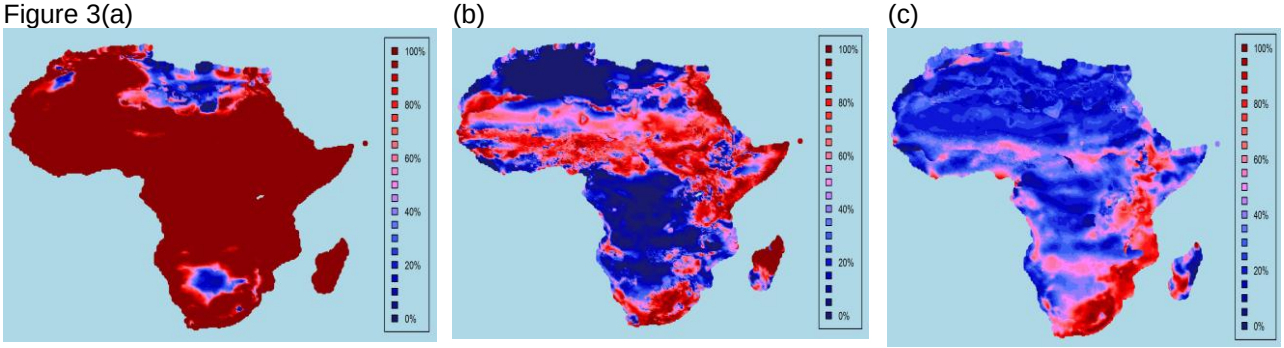
Table 2. Summary of the bioclimatic variables used within the final generalised additive models predicting the presence or absence of *Plasmodium falciparum,* Rift Valley Fever and *Babesia bigemina*. The super-indices in the population variables indicate: 1 = used within the *P. falciparum* model, 2 = used within the Rift Valley Fever model, 3 = used within the *B. bigemina* model.

Code	Variable definition
Bioclimatic variables	
bio1	Mean annual temp.
bio4	Temp. seasonality
bio8	Mean temp. of wettest quarter
bio12	Annual precip.
bio14	Precip. of driest month
bio15	Seasonality of precip.
bio18	Precip. of warmest quarter
bio19	Precip. of coldest quarter
Population variables	
Human pop$^{1;2}$	Human population density
Sheep pop^{2}	Sheep population density
Goat pop^{2}	Goat population density
Catbuf pop$^{2;3}$	Cattle/Buffalo population density

Table 3. Summary of the descriptive statistics for generalised additive models incorporating an expectation-maximisation algorithm technique to predict the presence or absence of *Plasmodium falciparum*, Rift Valley Fever and *Babesia bigemina* according to different prior probabilities (*P*i) of the likelihood of a detected absence of the pathogen actually being a presence.

Descriptive statistics for final models					
	Pi=*0.0* *model*	Pi=*0.25* *model*	Pi=*0.50* *model*	Pi=*0.75* *model*	*Surveillance adjusted* Pi *value*
Plasmodium falciparum					
Log-likelihood value	N/A	N/A	N/A	N/A	N/A
Adjusted R-squared value	64.0%	63.5%	62.2%	60.9%	58.8%
Rift Valley Fever					
Log-likelihood value	-17741889	-9843718	-5724748	-2618860	N/A
Adjusted R-squared value	40.3%	38.3%	36.6%	35.3%	37.8%
Babesia bigemina					
Log-likelihood value	-8830516	-4907646	-2838955	-1324010	N/A
Adjusted R-squared value	49.2%	42.2%	38.8%	36.4%	44.7%

Figure 3(a) (b) (c)



(d) (e) (f)

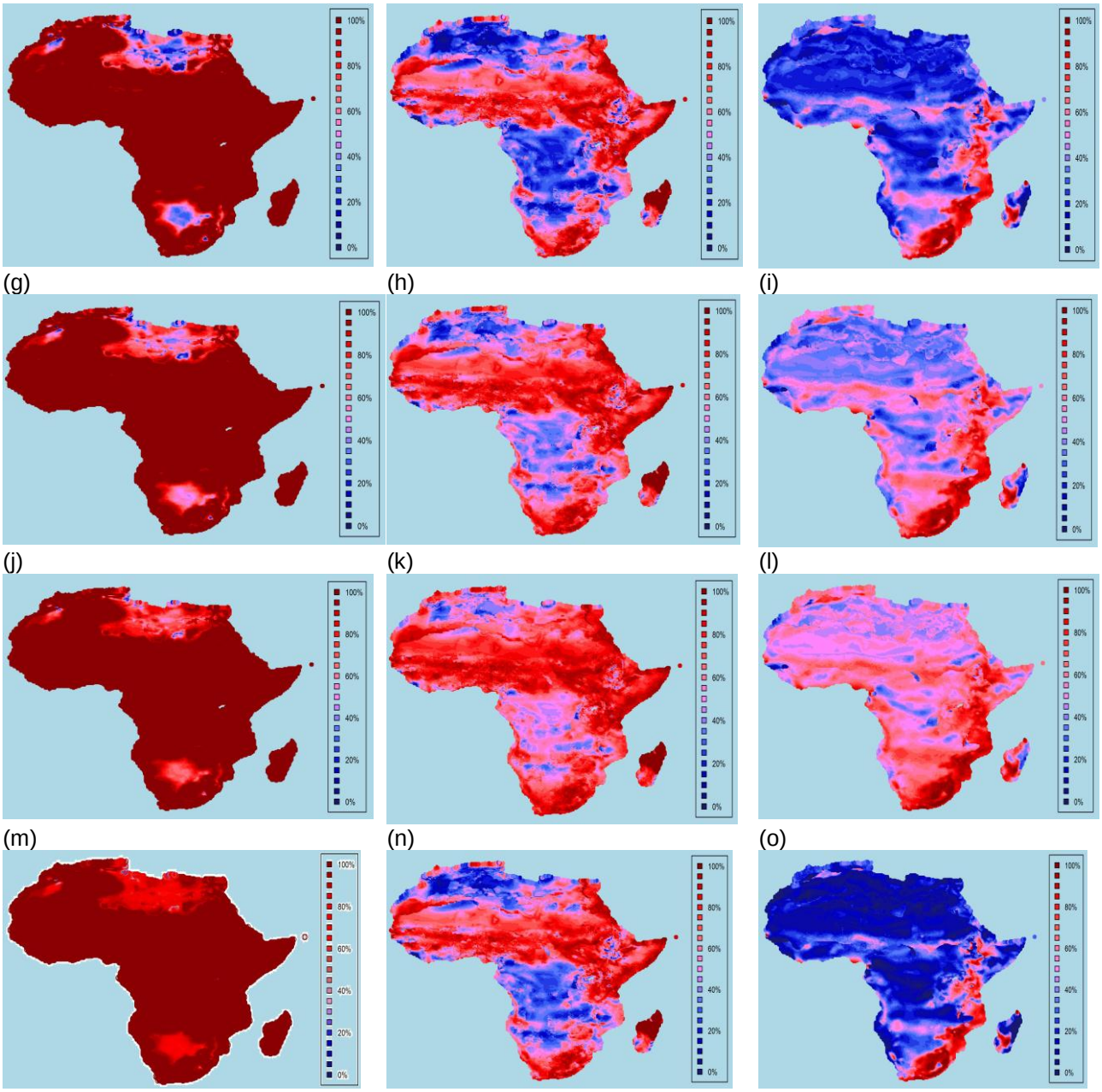


Figure 3. Presence of predicted pathogens using logistic regression modelling which incorporates an expectation-maximisation algorithm technique to predict presence according to different prior probabilities of the likelihood of a detected absence of a pathogen (when it has not been recorded in the EID2) actually being a presence. Figures 3a, 3d, 3g, 3j, 3m depict the predicted presence of *Plasmodium falciparum*, Figures 3b, 3e, 3h, 3k, 3n depict the predicted presence of Rift Valley Fever, and Figures 3c, 3f, 3i, 3l, 3o depict the predicted presence of *Babesia bigemina*, according to a prior probability (*Pi*) that a detected absence of the pathogen from within the EID2 database presence data is actually a presence with probability values of *Pi*=0.0 (Figures 3a-c), *Pi*=0.25 (Figures 3d-f), *Pi*=0.50 (Figures 3g-i), *Pi*=0.75 (Figures 3j-l) and surveillance adjusted *Pi* value (Figures 3m-o).

Discussion

The results of the EM-GAM modelling exercises suggest that the presence of pathogens can be predicted using this technique, however prior knowledge of the probability of reported absences of pathogens actually being presences would be useful in order to decide the best *Pi* values to incorporate into the model; the value of *Pi* cannot be estimated using the data, as previously suggested by Ward et al. (2009). Although our results show that models without any adjustments for the absence of a pathogen actually being a presence (*Pi*=0.0) had the highest log-likelihood values and were potentially therefore the best models, such

adjustments should be included as differences in surveillance and therefore reporting undoubtedly occur, as we have illustrated. It is worth noting that the predicted results of the models including surveillance adjusted *Pi* values which accounted for this issue differed little from the output of *Pi*=0.0 models. We therefore suggest that such an adjustment should be utilised as the best *Pi* estimate for EM-GAM modelling.

Future work should include testing for the accuracy of the outputs of EM-GAM modelling exercises such as these, by comparing outputs with test data for diseases/pathogens; either from previous modelling exercise outputs describing the presence of pathogens, or real-world data.

References

McIntyre, K.M., Setzkorn, C., Wardeh, M., Hepworth, P.J., Radford, A.D., Baylis, M., 2013. What, where and weather? Integrating open-source taxonomic, spatial and climatological information into a comprehensive database of livestock infections. In, Proceedings of the Annual Meeting, Society for Veterinary Epidemiology and Preventative Medicine, Madrid, Spain, 143-159. ISBN: 978-0-948073-20-5.

National Center for Biotechnology Information, 2012a. US National Library of Medicine, Bethesda, Maryland, US. The NCBI Medical Subject Headings (MeSH) database homepage. http://www.ncbi.nlm.nih.gov/mesh.

National Center for Biotechnology Information, 2012b. US National Library of Medicine, Bethesda, Maryland, US. The NCBI Nucleotide database homepage. http://www.ncbi.nlm.nih.gov/nuccore.

National Center for Biotechnology Information, 2012c. US National Library of Medicine, Bethesda, Maryland, US. PubMed. http://www.ncbi.nlm.nih.gov/pubmed/.