Grant agreement no. 243964

QWeCI

Quantifying Weather and Climate Impacts on Health in Developing Countries

M1.1.b - Completion of disease-climate relationships from pilot projects

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

Lead contractor : UNILIV
Coordinator of milestone : UNILIV
Evolution of milestone

Due date : M30
Date of first draft : 12 April 2013
Start of review : 20 April 2013
Milestone accepted : 25 April 2013

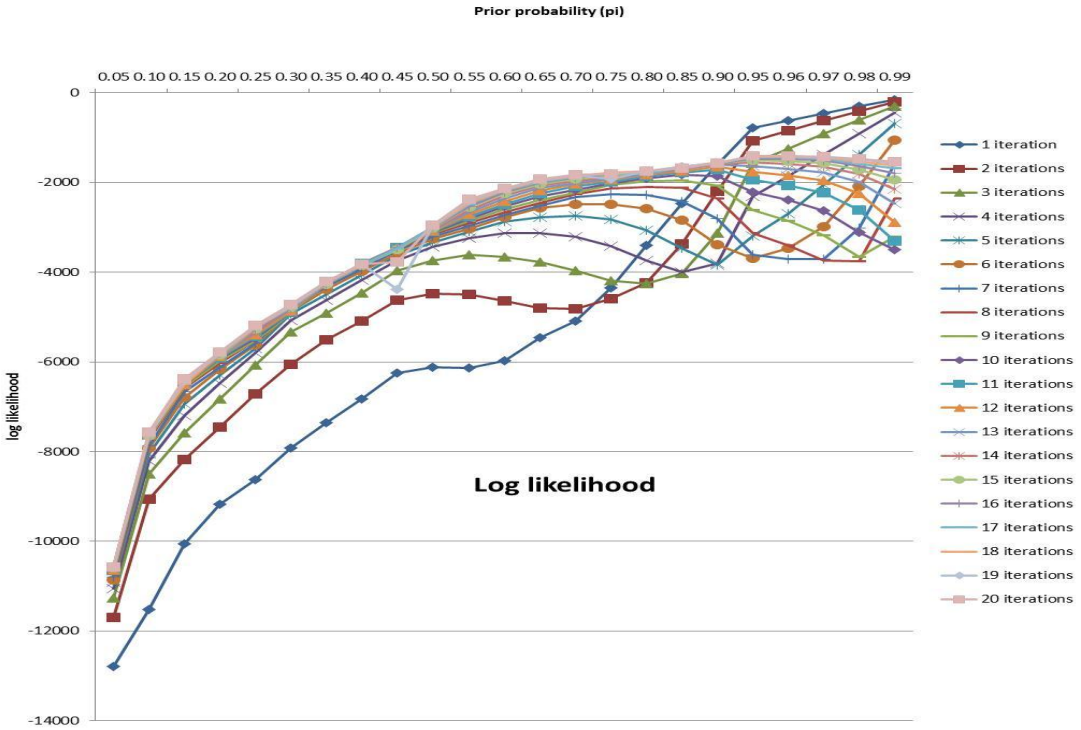
Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination Level		
PU	Public	PU
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)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

The capability to spatially map disease or pathogen data has been built into the ENHanCed Infectious Disease (EID2) database at the same spatial resolution as fields of climate data and some host population data. Methodologies incorporating the spatial distribution of pathogens using information obtained automatically from the literature and from nucleotide sequences submitted to the Genbank and other databases (accessible via the NCBI Taxonomy Database (http://www.ncbi.nlm.nih.gov/taxonomy) have also been developed (see McIntyre, in Proceedings of the Annual Meeting 2013, Society for Veterinary Epidemiology and Preventive Medicine, Madrid, Spain), and this information is available within the EID2 database. Both the presence-only and climate/population data can be exported for each pathogen from within the EID2.

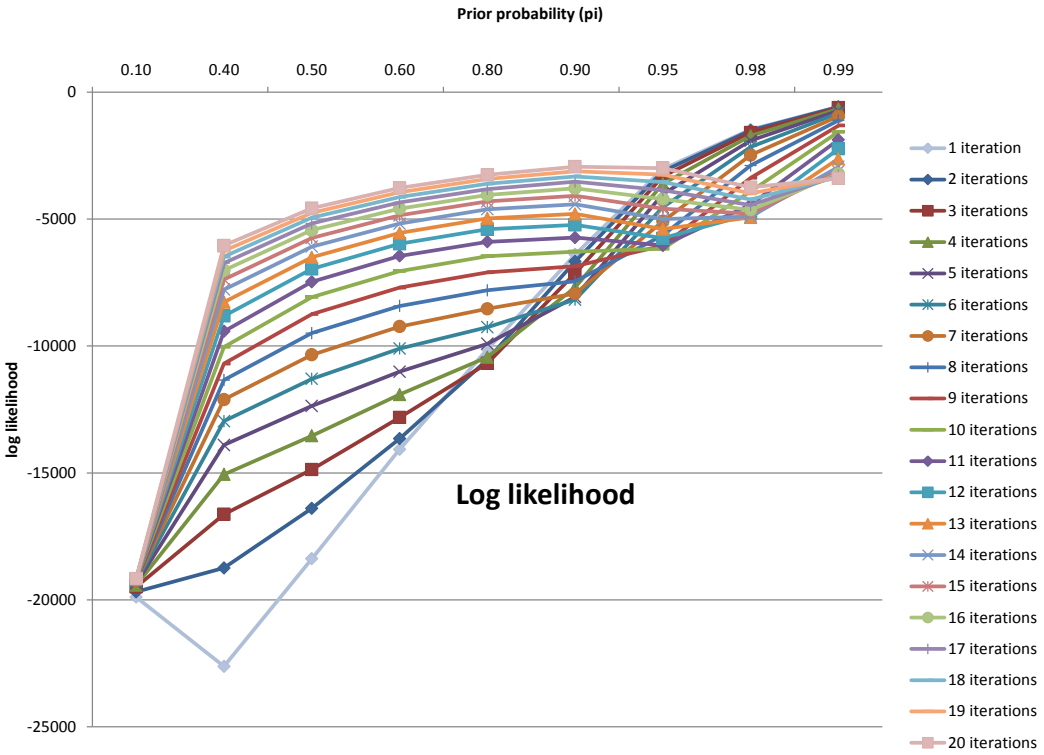
Code to run logistic regression modeling using an expectation-maximization (EM) algorithm technique (see Ward et al., 2009) has then been written for use within the R statistical package. This code uses data exported from the EID2 to predict the presence of pathogens given certain climate variables, according to different prior probabilities (*Pi*) of the likelihood of an absence of a pathogen actually being a presence. Initially, multiple models were tested in which different (a) numbers of iterations and (b) *Pi* values incorporated within the EM technique were used, 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 an absence actually being a presence could be predicted without prior knowledge (Figures 1a and b). 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.

Figure 1. Log likelihood values for logistic regression models of (a) *Plasmodium falciparum* and (b) Rift Valley fever which incorporate an expectation-maximization (EM) algorithm technique to predict presence according to different prior probabilities (*Pi*) of the likelihood of an absence actually being a presence. Values are depicted for models which tested different: numbers of iterations; and *Pi* values, to maximize the log likelihood value of the model.

(a) *Plasmodium falciparum*



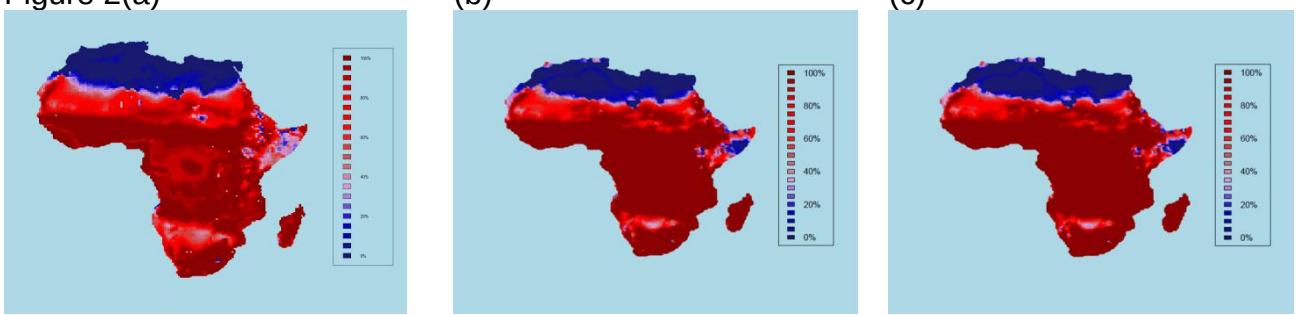
(b) Rift valley Fever



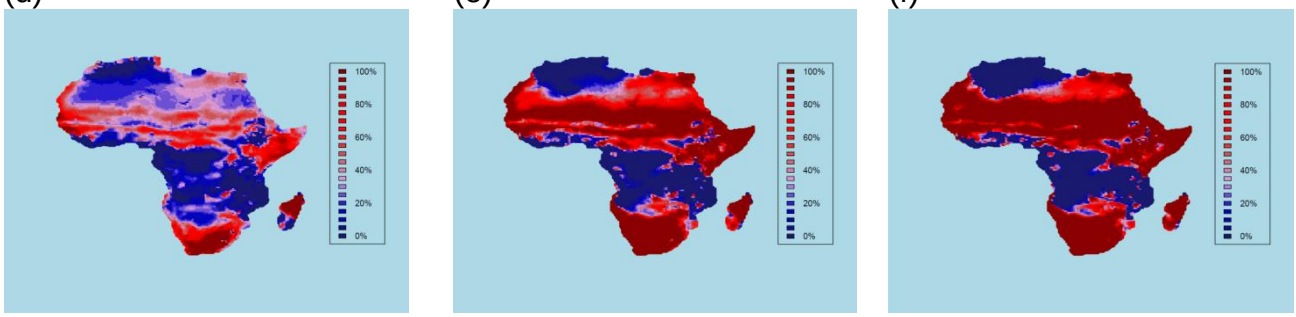
This modeling exercise was undertaken studying multiple pathogens which the QWeCI project aims to examine, but which have different extents of presence across the African continent. *Plasmodium falciparum* was examined as a pathogen with a ubiquitous presence, Rift Valley fever has a reasonably wide although spatially restricted presence, and *Babesia bigemina* has a very localised presence (Figures 2a-f).

Figure 2. Presence of predicted pathogens using logistic regression modelling which incorporates an expectation-maximisation (EM) algorithm technique to predict presence according to different prior probabilities of the likelihood of an absence actually being a presence. Figures 2a-c depict the predicted presence of *Plasmodium falciparum*, Figures 2d-f depict the predicted presence of Rift Valley fever, and Figures 2g-i depict the predicted presence of *Babesia bigemina*, according to a prior probability (*Pi*) that an absence of the pathogen from within the EID2 database presence data is actually a presence with probability values of *Pi*=0.1 (Figures 2a, 2d and 2g), *Pi*=0.5 (Figures 2b, 2e and 2h), and *Pi*=0.7 (Figures 2c, 2f and 2i).

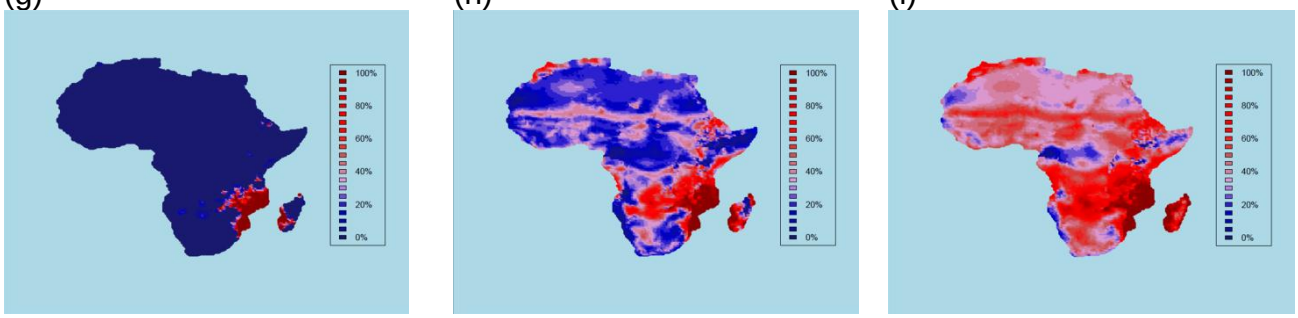
Figure 2(a) (b) (c)



(d) (e) (f)



(g) (h) (i)

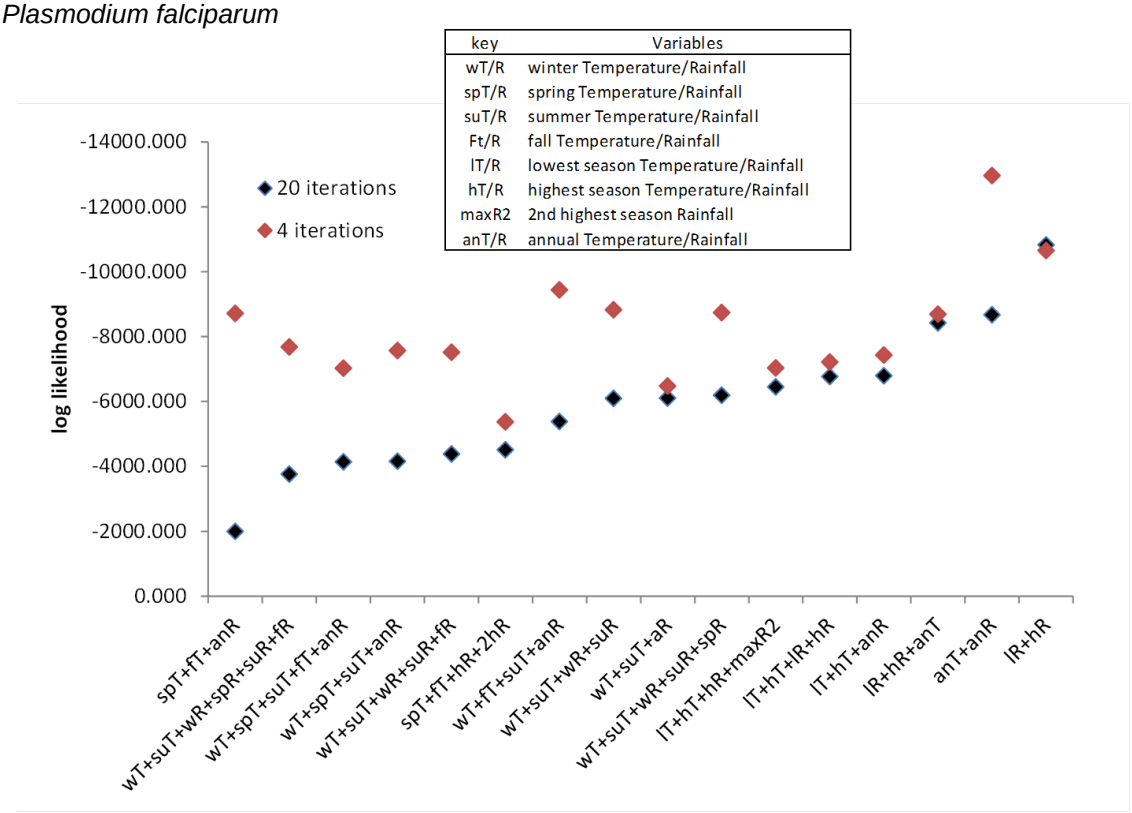


The results of these exercises suggested that the logistic regression models converged after 20 iterations and but also that the prior probability of a pathogen absence actually being a presence could not be predicted without some kind of (Bayesian) prior knowledge of the distribution of the pathogen, as previously suggested by Ward et al. (2009). Other approaches to ascertain the *Pi* value to use within modelling could incorporate location-specific priors using subject-matter knowledge of the locations concerned, for instance, describing how intensively pathogen presence has been investigated in an area.

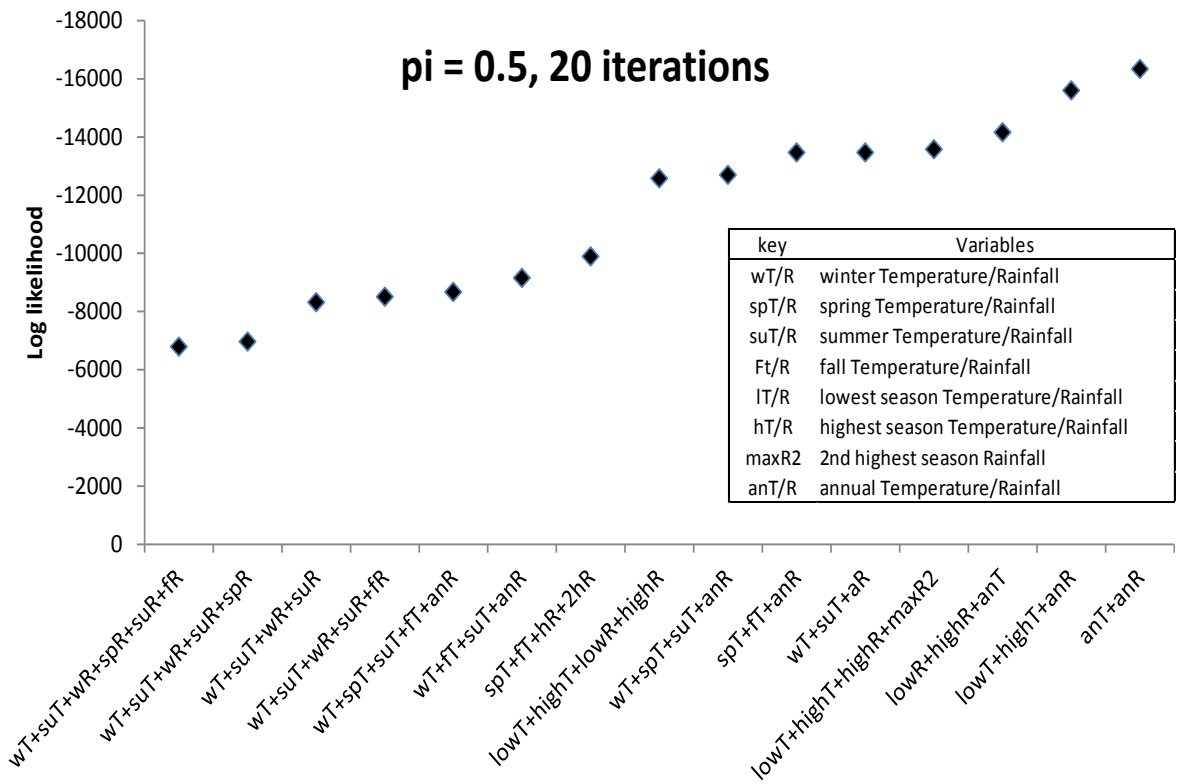
Aside from issues to decide *Pi* values, models in which different assemblages of climate variables were incorporated have been run for *Plasmodium falciparum*, Rift Valley Fever and *Babesia bigemina* (Figures 3a-c). Within these models, the objective is to ascertain the assemblage of climate variables which maximises the log likelihood value for the presence of the pathogen.

Figure 3. Log likelihood values for logistic regression models of (a) *Plasmodium falciparum*, (b) Rift Valley Fever, and (c) *Babesia bigemina* which incorporate an expectation-maximisation (EM) algorithm technique to predict the presence of the pathogen, according to a prior probability value of *Pi*=0.5 for the likelihood of an absence actually being a presence.

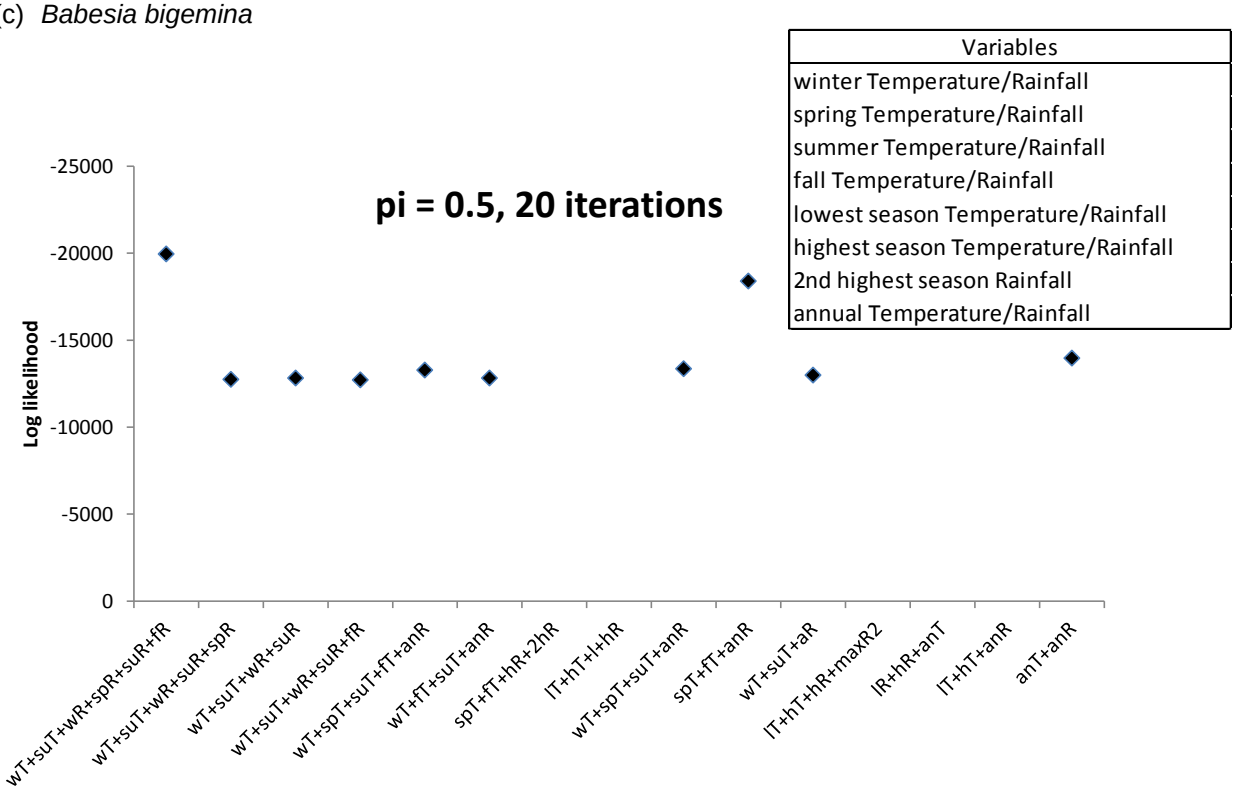
(a) *Plasmodium falciparum*



(b) Rift Valley Fever



(c) *Babesia bigemina*



Further work incorporating variables such as location-specific priors describing how intensively pathogen presence has been investigated in an area will be undertaken to try and ascertain the best *Pi* value to use within future modelling. In addition, further models will be run incorporating climate variables at other temporal resolutions (for example monthly) and also including host population density and potentially other land-use and landscape variables. Once best models have been isolated, their outputs will be compared to previous modelling exercises describing the presence of pathogens, or perhaps to real-world data.