



Using global maps to predict the risk of dengue in Europe[☆]



David J. Rogers^{a,*}, Jonathan E. Suk^{b,1}, Jan C. Semenza^{b,1}

^a University of Oxford, Department of Zoology, South Parks Road, Oxford OX1 3PS, United Kingdom

^b European Centre for Disease Prevention and Control, Tomtebodavägen 11A, 17183 Stockholm, Sweden

ARTICLE INFO

Article history:

Received 21 February 2013

Received in revised form 2 July 2013

Accepted 12 August 2013

Available online 21 August 2013

Keywords:

Dengue

Risk maps

Satellite data

Discriminant analysis

Mahalanobis distance

Fleiss kappa

ABSTRACT

This article attempts to quantify the risk to Europe of dengue, following the arrival and spread there of one of dengue's vector species *Aedes (Stegomyia) albopictus*. A global risk map for dengue is presented, based on a global database of the occurrence of this disease, derived from electronic literature searches. Remotely sensed satellite data (from NASA's MODIS series), interpolated meteorological data, predicted distribution maps of dengue's two main vector species, *Aedes aegypti* and *Aedes albopictus*, a digital elevation surface and human population density data were all used as potential predictor variables in a non-linear discriminant analysis modelling framework. One hundred bootstrap models were produced by randomly sub-sampling three different training sets for dengue fever, severe dengue (i.e. dengue haemorrhagic fever, DHF) and all-dengue, and output predictions were averaged to produce a single global risk map for each type of dengue. This paper concentrates on the all-dengue models. Key predictor variables were various thermal data layers, including both day- and night-time Land Surface Temperature, human population density, and a variety of rainfall variables. The relative importance of each may be shown visually using rainbow files and quantitatively using a ranking system. Vegetation Index variables (a common proxy for humidity or saturation deficit) were rarely chosen in the models. The kappa index of agreement indicated an excellent (dengue haemorrhagic fever, Cohen's kappa = 0.79 ± 0.028 , AUC = 0.96 ± 0.007) or good fit of the top ten models in each series to the data (Cohen's kappa = 0.73 ± 0.018 , AUC = 0.94 ± 0.007 for dengue fever and 0.74 ± 0.017 , AUC = 0.95 ± 0.005 for all dengue). The global risk map predicts widespread dengue risk in SE Asia and India, in Central America and parts of coastal South America, but in relatively few regions of Africa. In many cases these are less extensive predictions than those of other published dengue risk maps and arise because of the key importance of high human population density for the all-dengue risk maps produced here. Three published dengue risk maps are compared using the Fleiss kappa index, and are shown to have only fair agreement globally (Fleiss kappa = 0.377). Regionally the maps show greater (but still only moderate) agreement in SE Asia (Fleiss kappa = 0.566), fair agreement in the Americas (Fleiss kappa = 0.325) and only slight agreement in Africa (Fleiss kappa = 0.095). The global dengue risk maps show that very few areas of rural Europe are presently suitable for dengue, but several major cities appear to be at some degree of risk, probably due to a combination of thermal conditions and high human population density, the top two variables in many models. Mahalanobis distance images were produced of Europe and the southern United States showing the distance in environmental rather than geographical space of each site from any site where dengue currently occurs. Parts of Europe are quite similar in Mahalanobis distance terms to parts of the southern United States, where dengue occurred in the recent past and which remain environmentally suitable for it. High standards of living rather than a changed environmental suitability keep dengue out of the USA. The threat of dengue to Europe at present is considered to be low but sufficiently uncertain to warrant monitoring in those areas of greatest predicted environmental suitability, especially in northern Italy and parts of Austria, Slovenia and Croatia, Bosnia and Herzegovina, Serbia and Montenegro, Albania, Greece, south-eastern France, Germany and Switzerland, and in smaller regions elsewhere.

© 2013 The Authors. Published by Elsevier B.V. All rights reserved.

[☆] This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-No Derivative Works License, which permits non-commercial use, distribution, and reproduction in any medium, provided the original author and source are credited.

* Corresponding author. Present address: Hilltop Cottage, Horton-cum-Studley, Oxford OX33 1AU, United Kingdom. Tel.: +44 01865 351348; mobile: +44 07762260802. E-mail addresses: david.rogers@zoo.ox.ac.uk (D.J. Rogers), jonathan.suk@ecdc.europa.eu (J.E. Suk), jan.semenza@ecdc.europa.eu (J.C. Semenza).

¹ Tel.: +46 08 5860 1000.

1. Introduction

Approximately one third of the world's human population are exposed to the risk of dengue (Rogers et al., 2006). Each year there are between 50 and 100 million recorded cases of dengue fever, 500,000 cases of severe dengue (usually referred to as dengue haemorrhagic fever, DHF) requiring hospitalisation and between 20,000 and 25,000 deaths, mostly of children (Gubler, 2006). Although today dengue is mostly a tropical disease, historically it occurred as far North as Philadelphia in the USA and, less than 100 years' ago, caused 1000 deaths amongst one million infected people in Greece (Reiter, 2010). Development of public health services, environmental sanitation and improvement of construction methods have removed dengue from most temperate areas, but increased international trade and tourism currently increase the risk of introduction or re-introduction of one or more of dengue's vectors, and of the disease itself in travellers returning from dengue infected areas. Whilst dengue's major vector species, *Aedes aegypti*, which once had a range that extended into temperate Europe (Reiter, 2010), is now restricted mostly to the tropics, the alternative vector species *Aedes (Stegomyia) albopictus* – the Asian tiger mosquito – now occurs on all the world's populated continents, with relatively recent successful introductions into Europe. The global distribution and spread of this vector has been analysed in a number of recent publications (Benedict et al., 2007; Tatem et al., 2006a,b), including in the twin contexts of the environmental suitability of the destination areas and the volume of international (mostly sea) traffic going to those areas. Concern over the arrival of *Aedes albopictus* in Europe was shown to be justified by the later arrival and transmission by it in Europe of both the alphavirus chikungunya, following a large outbreak of this disease in the Indian Ocean (Pialoux et al., 2007), and dengue in mainland France (La Ruche et al., 2010) and Croatia (Gjenero-Margan et al., 2011). An outbreak of dengue in Madeira in late 2012 (transmitted by *A. aegypti*) involved over 2000 cases, the majority residents of the island, but with a minority of cases in tourist visitors to the island, diagnosed after they had returned to their home countries in the rest of Europe (http://ecdc.europa.eu/en/press/news/Lists/News/ECDC_DispForm.aspx?List=32e43ee8%2D230%2D4424%2Da783%2D85742124029a&ID=809).

Another trend that may favour the spread of vectors and diseases is climate and environmental change. Much has been written on the impacts of climate change on vector-borne diseases, but careful analysis suggests that many examples of such changes can also be explained by alternative causes, from better reporting (e.g. in the case of Lyme disease in some parts of Europe) to drug resistance (e.g. malaria in East Africa) (Rogers et al., 2006). Several examples of the arrival establishment and spread of a number of vector-borne diseases, and what we may learn from them, are reviewed by (Randolph and Rogers, 2010). Undoubtedly climate change and other environmental changes will eventually affect vector-borne and other diseases, if such changes are large enough to affect important demographic rates and transmission processes. Vector-borne diseases are particularly sensitive to climate and it is therefore among this group of diseases that we might expect to detect first the impacts of climate change. Whether such diseases will increase or decrease in distribution and severity depends on the relative impacts of the climatic changes on factors that increase or decrease transmission.

The present work arose from the desire of the European Centre for Disease Prevention and Control (ECDC) to assess the possibility of dengue establishing within Europe, or on its fringes. Currently, most dengue cases reported in Europe are of travellers returning with infections contracted in exotic destinations, mostly India, Thailand, Indonesia, Mexico and Brazil. A few cases have also been

observed in patients from EU overseas territories. Bearing in mind the recent spread of *A. albopictus* in Europe, and of disease events attributable to this species, the potential for the re-establishment of dengue in Europe appeared to warrant further study.

The objective of this study was therefore to obtain a better understanding of the various factors that contribute to the determination of the risk of dengue in continental Europe. This was achieved by developing global risk maps for dengue, based on and extending previous work on this topic (Rogers et al., 2006), and examining the predictions of such maps for Europe.

A variety of global predictions now exist for dengue (Degallier et al., 2010; Hales et al., 2002; Jetten and Focks, 1997; Patz et al., 1998; Rogers et al., 2006; Simmons et al., 2012; Bhatt et al., 2013), developed from different databases and using different modelling approaches. Model outputs are rarely compared to see whether or not there is any consensus in areas predicted to be at risk of this disease. Here we compare three different predictions, based on non-linear discriminant analysis (this study), logistic regression (Hales et al., 2002) or boosted regression tree (Bhatt et al., 2013) modelling approaches. Considerable differences are found between pairs of models, and the consensus of all three models is classed as only 'fair' (Landis and Koch, 1977), with disagreement around the edges of dengue's predicted distributions, or in Africa where dengue is considered to be under-reported.

2. Materials and methods

2.1. Disease and vector data

The search strategy to produce the disease and vector database was an extended version of that described elsewhere (Rogers et al., 2006). Briefly, searches were made in PubMed, Web of Science and Promed databases using the search terms '*Aedes aegypti*', '*Aedes albopictus*', 'dengue fever' and 'dengue haemorrhagic fever'. Searches were restricted to publications that appeared between 1960 and the end of 2009, the retrieval of which was considerably easier (electronically or by direct library searches) than of articles published before this time. All search Abstracts were read and all papers were acquired and read in full if the Abstracts suggested they might contain geographical information about the vectors or disease. Data were extracted from these into an electronic database, and place names were geolocated using gazetteers, including the GEOnet Names Server (GNS) of the National Geospatial-Intelligence Agency's (NGA) and the U.S. Board on Geographic Names' (US BGN) database of foreign geographic feature names (<http://earth-info.nga.mil/gns/html/>, last accessed December 2012), Encarta, (Microsoft Corporation, WA, USA), the Alexandria Digital Library Gazetteer (URL: <http://middleware.alexandria.ucsb.edu/client/gaz/adl/index.jsp>, last accessed November 2012) and the Getty Thesaurus of geographic names online, (URL: http://www.getty.edu/research/conducting_research/vocabularies/tgn/index.html, last accessed December 2012).

An improved version of the dengue database used in this study has recently been developed by (Brady et al., 2012) as part of a project to estimate the global burden of dengue (Bhatt et al., 2013).

One common feature of disease and vector data is that location names frequently refer to polygons (usually administrative units) rather than to precise points. This information was also stored in the database, and administrative unit geolocations (i.e. of polygon centres) were later extracted from the SALB (<http://www.unsalb.org/>, last accessed November 2012) or GAUL (<http://www.fao.org/giews/english/shortnews/sdrngiewsgaul.htm>, last accessed November 2012) databases. Totals of 2736 unique dengue fever and 736 unique dengue haemorrhagic fever records were recorded in the database (points and polygons, resolved to

the 1/15th degree resolution of all the models). When combined, there was a total of 2927 unique records at this resolution for any type of dengue. The geographic distributions of the data points for dengue fever, dengue haemorrhagic fever and all-dengue are shown in Supplementary Information Figure 1 and the reporting decades for all records in the database are given in Supplementary Information Table 1 which shows that more than half of the data points were gathered in the first decade of this century.

Most disease databases record only the presence of the diseases in question, rarely their absence. Presence points come from within the accessible (to the disease) realised niche (Barve et al., 2011; Soberon and Peterson, 2005) and may therefore result in model predictions which over-estimate the inhabited range of the disease (because inaccessible areas may be deemed suitable biotically and abiotically). Although some distribution modelling software claims to be able to use presence-only data, in practice other locations chosen at random throughout the area are often taken effectively as absence data (e.g. the back-ground data in MAXENT; Phillips et al., 2004). In the present study, series of pseudo-absence points were chosen in a two-step process. In the first step, they were selected at random at distances of no less than 0.5 degrees and no greater than 5 degrees away from any of the presence data points. A single model (described below) was then run for each entire set of presence and pseudo-absence data (using a maximum of ten predictor variables selected in ways described below), from which was generated a Mahalanobis distance image of environmental distance of every pixel from any of the presence pixel centroids (Green, 1978; Krzanowski and Marriott, 1995). The Mahalanobis distance (D^2) is a covariance adjusted measure of difference between sets of environmental conditions; small values indicate similar conditions and large values dissimilar ones. It is calculated as follows:

$$D_{12}^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{C}_w^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = \mathbf{d}' \mathbf{C}_w^{-1} \mathbf{d} \quad (1)$$

where the subscripts refer to groups 1 (e.g. for vector or disease absence) and 2 (e.g. for vector or disease presence), $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ refer to the vectors of the means of the variables of the two groups (the 'centroids'), \mathbf{C}_w^{-1} is the inverse of the within-groups covariance (=dispersion) matrix, $\mathbf{d} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ and the symbol (') indicates a matrix or vector transpose.

By examining the frequency distribution of Mahalanobis distances of known presence points correctly identified as such by the model, and of suspected pseudo-absence points, it was possible to select a Mahalanobis distance threshold that reduced to a minimum the likely misclassification errors, and this threshold was used in the second step of pseudo-absence point generation when both geographic and Mahalanobis distance criteria were used together. Pseudo-absence points were then chosen to be no closer to any known presence point than the selected thresholds, in both geographic and environmental spaces, and again never more than 5 degrees away from any of them. The Mahalanobis distance thresholds chosen were 7, 5 and 6, and totals of 14,000, 4000 and 15,000 pseudo-absence points were generated respectively for the dengue fever, dengue haemorrhagic fever and all-dengue models. Each series of presence and pseudo-absence points is referred to as a 'training set' below.

2.2. Environmental and other data

Time series of nominal 1 km spatial resolution MODIS v4 data from the NASA Terra satellite were downloaded from NASA's EOS data gateway (<http://ladsweb.nascom.nasa.gov/data/search.html>, last accessed December 2012) for five complete years, January 2001 to December 2005. Further details of this imagery are given in (Scharlemann et al., 2008). Table 1 lists the MODIS and other derived products used in this study.

Table 1

NASA MODIS and other multi-temporal data used in the present study. 'Number' indicates the identifier number used in the variable filenames in the other tables.

MODIS data	Number	Name
MIR	3	Middle Infra-red
dLST	7	Day-time Land Surface Temperature
nLST	8	Night-time Land Surface Temperature
NDVI	14	Normalised Difference Vegetation Index
EVI	15	Enhanced Vegetation Index
CMORPH	50	CMORPH satellite-derived rainfall estimate
WORLDCLIM	57	WORLDCLIM interpolated rainfall estimate

MODIS data do not contain any useful proxies for rainfall, an important determinant for many vector-borne diseases. This study therefore supplemented the MODIS data with two rainfall estimates:

- Satellite-derived rainfall estimates from the CMORPH project (Joyce et al., 2004) provided at a resolution of 0.25 degrees (http://www.cpc.noaa.gov/products/janowiak/cmorph_description.html, last accessed December 2012).
- Meteorological station derived WORLDCLIM climate surfaces at 1/120th degree resolution (<http://www.worldclim.org/>, last accessed December 2012) (Hijmans et al., 2005). Obviously obtaining this fine resolution imagery from meteorological station records required fairly sophisticated interpolation of the original point data; the WORLDCLIM project checked the accuracy of interpolation by missing out certain stations and using the remainder to construct climate surfaces, from which the data for the omitted point could be read, and checked against the correct readings. The authors claim cross validation (i.e. 'miss one out') accuracy of <10 mm rainfall per month 'in the vast majority of places' (Hijmans et al., 2005, p. 1969), although accuracy was very much worse when more stations were missed out.

The CMORPH data span only the latitudes 60°N to 60°S and are available as monthly composites, whilst the WORLDCLIM data cover the globe from 90°N to 60°S and are available as synoptic monthly values.

In addition to environmental data, the models used an image of estimated human population density per square km, derived from the GRUMP project (<http://sedac.ciesin.columbia.edu/data/collection/grump-v1/sets/browse>, last accessed December 2012) (Balk et al., 2006) and a 1 km resolution digital elevation surface produced by the MODIS team (now updated to 250 m resolution, https://lpdaac.usgs.gov/products/modis_products_table/mod44w, last accessed December 2012).

2.3. Environmental data processing

All multi-temporal environmental data were temporal Fourier processed, a transformation that produces descriptors of environmental signals that are both robust statistically and meaningful biologically (Rogers, 2000, 2006; Rogers and Williams, 1993, 1994). These descriptors for each channel were the overall mean signal, the amplitudes and phases of the annual, bi-annual and tri-annual cycles of activity, the maximum, minimum and variance (i.e. ten Fourier variables per channel). Full details of MODIS data processing are given in Scharlemann et al. (2008). All environmental data were temporal Fourier processed at their original resolution and the resulting imagery was either progressively aggregated (by averaging) to 1/15th degree resolution (all MODIS and WORLDCLIM data) or cubic spline-interpolated to 1/15th degree resolutions (CMORPH data). Similarly, the human population density surface was aggregated to 1/15th degree resolution. Global models for the current distribution of dengue and its vectors were then developed at this

resolution. For each of the training set databases all of the satellite and other data were extracted for each of the unique presence and pseudo-absence points.

2.4. Distribution modelling

Spatial analysis in general is reviewed by Pfeiffer et al. (2008), in which the majority of techniques described are illustrated using the same epizootiological dataset, of bovine tuberculosis in Great Britain from 1986 to 1999. Species distribution models are reviewed, amongst others, by Elith et al. (2006, 2008) who concluded that many of the newer methods of predicting species' distributions (e.g. MAXENT, Boosted Regression Trees) are better than the more traditional methods such as logistic regression, Generalised Additive Models (GAMs) or the widely used GARP (Stockwell and Peters, 1999).

As various authors have pointed out, each modelling approach sits on a continuum between statistical and biological models (Rogers, 2006) or between data-driven and knowledge-driven approaches (Stevens and Pfeiffer, 2011). Virtually none of the currently available models is a true spatial model; in making predictions they do not consider either the occurrence of the target species in adjacent spatial units (pixels), or the environmental conditions in those pixels (Rogers and Sedda, 2012), both of which would be considered by properly geospatial models.

The choice for a young biologist is effectively to choose between species' distribution models that describe patterns well and those that concentrate more on the underlying biological processes that determine species' distributions. In the authors' view, variants of discriminant analysis (not included in Elith et al., 2006) strike a realistic compromise between these two approaches. Discriminant analysis has a central assumption of a multi-variate normal response of species to environmental variables, an assumption that has pervaded much of ecological theory from the time of MacArthur (1972). Non-linear discriminant analysis (NLDA) can cope with cases where overall responses are not multivariate normal. It does so by clustering, before modelling, the environmental data for species' presence and absence into two series of multi-variate normal clusters. NLDA then generates a non-linear discriminant function that separates presence and absence clusters in environmental space. These functions are then used to assign each pixel to the presence or absence clusters; the output may be either a binary presence/absence prediction or, more usefully, the posterior probability with which each pixel belongs to the presence or absence clusters. NLDA was implemented for the present study in the ways described in Rogers (2006). Briefly, each entire dataset was first clustered using the *k-means cluster* algorithm in SPSS v. 18 (WIBM) into 3 (dengue haemorrhagic fever) or 5 (dengue fever and all-dengue fever) clusters each for the presence and pseudo-absence points. 100 bootstrap samples were taken from each training set, with equal numbers of presence and absence points sampled, with replacement, and ensuring a minimum sample size of 30 for any particular cluster. Using equal total numbers of presence and absence points ensures greater model accuracy (McPherson et al., 2004), and sampling with replacement makes it more likely that the relationship of the bootstrap sample to the training set is similar to that between the training set and the real world of which it is a sample. The variability across the set of bootstrap samples – both in terms of the variables selected by the model and the resulting model predictions – should therefore indicate how likely it is that the training set sample adequately captures the full range of conditions inhabited by the species being modelled.

Within each bootstrap model, ten variables from the entire set of available predictors were selected in a forward step-wise inclusive fashion, on the basis of minimising the corrected Akaike's Information Criterion (AICc) (Burnham and Anderson, 2002). Text files

were kept of the results of each bootstrap model, including the selected variables, their mean values per cluster, their correlation matrix and the model accuracy in terms of a wider range of accuracy statistics (kappa, sensitivity, specificity and Area Under the Curve, AUC), all described in Table 1 of Rogers (2006), and the True Skill Statistic (Allouche et al., 2006). Each accuracy statistic may be criticised (for example see Lobo et al., 2008 for a discussion of the AUC) and some should not be used on their own (for example, sensitivity is meaningless without also considering specificity) but the full range is reported here for broad comparison with other models of dengue. The AICc used in the present models does not have any obvious drawbacks and also allowed the correct number of predictor variables (up to a maximum of ten) to be chosen for the output bootstrap maps, because the AICc is penalised as more variables are added (Burnham and Anderson, 2002). In the event, most models used all ten variables.

For each bootstrap model an output image file – a risk map – was produced using the model results applied to the selected input data layers for that particular model. Output was expressed in terms of posterior probabilities on the scale of 0 to 1 (Rogers, 2006). Discriminant analytical models calculate the position of each point on the multi-variate surfaces determined by the input data. The underlying assumption of multi-variate normality here ensures that the outputs are real probabilities in the statistical sense (i.e. they are the probabilities that each point 'belongs to' each cluster in the model, on the basis of the similarity of its environmental conditions to those of each cluster). Other modelling approaches produce results that are also usually expressed on the scale of zero to 1.0 (logistic regression, random forest, GARP and other models) but these are not strictly speaking probabilities (for example, random forest and GARP model outputs are proportional frequencies, not statistical probabilities).

The 100 risk maps produced from the 100 bootstrap samples were averaged to produce the output prediction – the final risk map for each model. The text output files from the 100 bootstrap models were also combined in several ways to produce a list of the ranking of each predictor variable, a list of accuracy statistics and a 'rainbow file' that allows much of this information to be appreciated visually (Rogers, 2006). Each row in the rainbow file represents 1 bootstrap model and each column represents 1 of the predictor variables made available to all the models. The rows are arranged in model order from best (the first row) to worst (the last row). Within each row, the variable selected first in that model is coloured red, the one selected second is coloured orange, and so on across the rainbow colour scale (red, orange, yellow, green, blue, indigo, etc.). Variables not selected are left a neutral background colour. A variable selected first across all the models therefore would show up as a single column of solid red in the rainbow file. Solid columns of single colours indicate that the same variables are being chosen across all models, in a sequence order determined by their rainbow colours. On the other hand, a rainbow file that looks much less organised, with patches of different colours in different places, arises when the models differ greatly in the variables selected, and the order in which those variables are chosen.

Risk maps were produced as described in this section for the training set databases of the two widespread mosquito vectors of dengue, *A. aegypti* and *A. albopictus*, and those of dengue fever, dengue haemorrhagic fever and of both dengue fever and DHF combined (referred to as the 'all-dengue' map). Two versions of the latter were produced, one using just the environmental, digital elevation and human population density data and one which, in addition to these, also included in the available predictor variable dataset the predicted global distribution maps of both vector mosquito species (these two maps were produced from the mosquito training set data by following exactly the same modelling process as described here for dengue). Inclusion of a variable in the

potential predictor dataset does not guarantee its inclusion in the step-wise selection process of model building, but it was felt to be of interest to see how often such vector maps were selected in the models. The results described in this paper concentrate on the all-dengue models that included the two mosquito species as potential predictor layers; comments are made when the conclusions drawn from the other two dengue maps differed considerably from those of the all-dengue map. The much greater number of database points for dengue fever than for dengue haemorrhagic fever means that the all-dengue model results show more similarity to those of dengue fever than of dengue haemorrhagic fever.

2.5. Assessing variable importance

For the routine bootstrap sample models, variable importance was assessed by calculating the mean rank of the selected variables across all the models in which they were made available. In each model, variables were assigned a rank of 1 to 10 if they were selected (1 = first variable selected; 2 = second variable, etc.) and a rank of 11 if not selected. Averaging ranks across all 100 bootstrap models established the mean rank importance of the variables in determining model predictions.

There is much discussion in the literature on species' distribution modelling on how to assess the contribution of each of the selected variables to model predictions. Whilst the first-selected of a whole group of variables is likely to be the single most important variable, the second and subsequent variables are operating in an environment already informed by one or more variables. Correlation between variables makes assessing the contribution of each rather difficult. Whilst Fourier analysis (that produces orthogonal, i.e. uncorrelated variables from time series data) reduces the likelihood of correlated variables within the predictor variable dataset, and step-wise variable selection methods reduce the chance of selecting variables that are strongly correlated with each other (because a step-wise method maximising any fitting criterion is unlikely to choose new variables strongly correlated with those already in the selected set; such variables are less likely to improve fit compared to other, less-correlated variables), the problem of assessing variable importance remains. Indeed the anti-correlation of variables selected in a stepwise fashion means that biologically important variables strongly correlated with the first selected variable are unlikely to be selected at all. Burnham and Anderson (2002) suggest a novel approach to this problem that has been applied to modelling Lassa fever distribution in West Africa (Fichet-Calvet and Rogers, 2009). The whole set of available predictor variables is sampled at random to generate many hundreds or thousands of combinations of variables that are then applied to the training sample data, generating one value of the AICc for each model from each random set of variables. Care is taken to include each variable the same number of times as all other variables across all combinations. Once completed, the total AICc for all models including each variable in turn is calculated. The variable with the lowest total AICc is the 'best' overall because, regardless of the other variables with which it was associated, it occurred in models that, on average, fitted the data best. The variable with the second to lowest total AICc is the next best variable, and so on. This combination approach establishes the importance of each variable in turn, in the same way that a series of races establishes the rank order of sprinters. As has been pointed out elsewhere, however, fitting variables to a species' distribution involves team-work; the top eleven sprinters are likely to be less good at a team task than a soccer team in which each member plays a rather different, and complementary role (Fichet-Calvet and Rogers, 2009). Thus we can identify the potential importance of variables individually by this method, but we should not expect the 'best ten' often to occur together in the models involving step-wise selection of variables.

In the present instance, 740 combinations of ten variables each were made by distributing each of the 74 available variables 100 times at random across the 740 combinations. All 740 combinations were then applied to each of the 100 bootstrap samples that had been used to generate the dengue prediction maps by non-linear discriminant analysis; a total of 74,000 models. The AICc was calculated for each model and the results analysed as suggested above.

2.6. Comparison of published dengue maps

Through direct contact with the original authors, two published dengue risk maps were obtained for comparison with those produced by the present study, to establish the degree of consensus between predictions. An early dengue risk maps used logistic regression methods applied to a map of mostly country-level dengue reports (sub-national where available) (Hales et al., 2002). A more recent map applied a boosted regression tree method to a point and polygon database similar to the one used in the present study (Bhatt et al., 2013). All model outputs were expressed on the scale of 0–1 and, even though the interpretation of these values is different, they were treated as similar in one respect; values lower than 0.5 were taken to represent predictions of disease absence and those equal to or greater than 0.5 predictions of disease presence. Predictions of the two higher spatial resolution maps were first averaged to the coarser scale of half of one degree of the lowest resolution map (Hales et al., 2002). Pair-wise comparisons of maps were made using Cohen's kappa index of agreement (Congalton, 1991). All three maps were compared using Fleiss's kappa which allows comparisons of more than two raters (=maps here) (Fleiss, 1971); this comparison was done both globally and by three regions; the Americas (longitudes 180W to 30W), Africa and Europe (longitudes 30W to 60E) and India and SE Asia (longitudes 60E to 180E). Both the Cohen and Fleiss indices scale from –1 to +1 where –1 indicates complete disagreement, +1 indicates complete agreement and a value of zero indicates a level of agreement that could be attributed to chance. Landis and Koch suggest the following ranges of agreement for Cohen's kappa: poor, $\kappa < 0.4$; good, $0.4 < \kappa < 0.75$; and excellent, $\kappa > 0.75$. They also suggest the following ranges for Fleiss's kappa: poor, $\kappa < 0$; slight, $0.01 < \kappa < 0.2$; fair, $0.21 < \kappa < 0.4$; moderate, $0.41 < \kappa < 0.6$; substantial, $0.61 < \kappa < 0.8$; and almost perfect, $0.81 < \kappa \leq 1.0$ (Landis and Koch, 1977).

2.7. Calculating environmental distances

The likelihood of any vector or disease spreading in a new area is at least partly dependent upon the environmental conditions in that area. Maps of the Mahalanobis distance, employed to define pseudo-absence points (Section 2.1), may be used as an indication of the suitability of areas for invasion, on the basis of their closeness in multi-variate space to conditions defining vector or disease presence. Each of the 100 bootstrap sample models also generated an output image of the Mahalanobis distance of each pixel to the environmentally nearest presence cluster. The 100 maps were again averaged to produce a single output Mahalanobis distance map that could be examined to establish the suitability of geographically marginal areas, such as Europe, for dengue.

3. Results

Fig. 1 shows the global all-dengue risk map and Table 2 lists the mean values of the environmental predictor variables for the top model (lowest AICc value) in the 100 bootstrap models contributing to it. Each bootstrap result is different from all the others, but the example in Table 2 is typical of the top bootstrap models in terms of

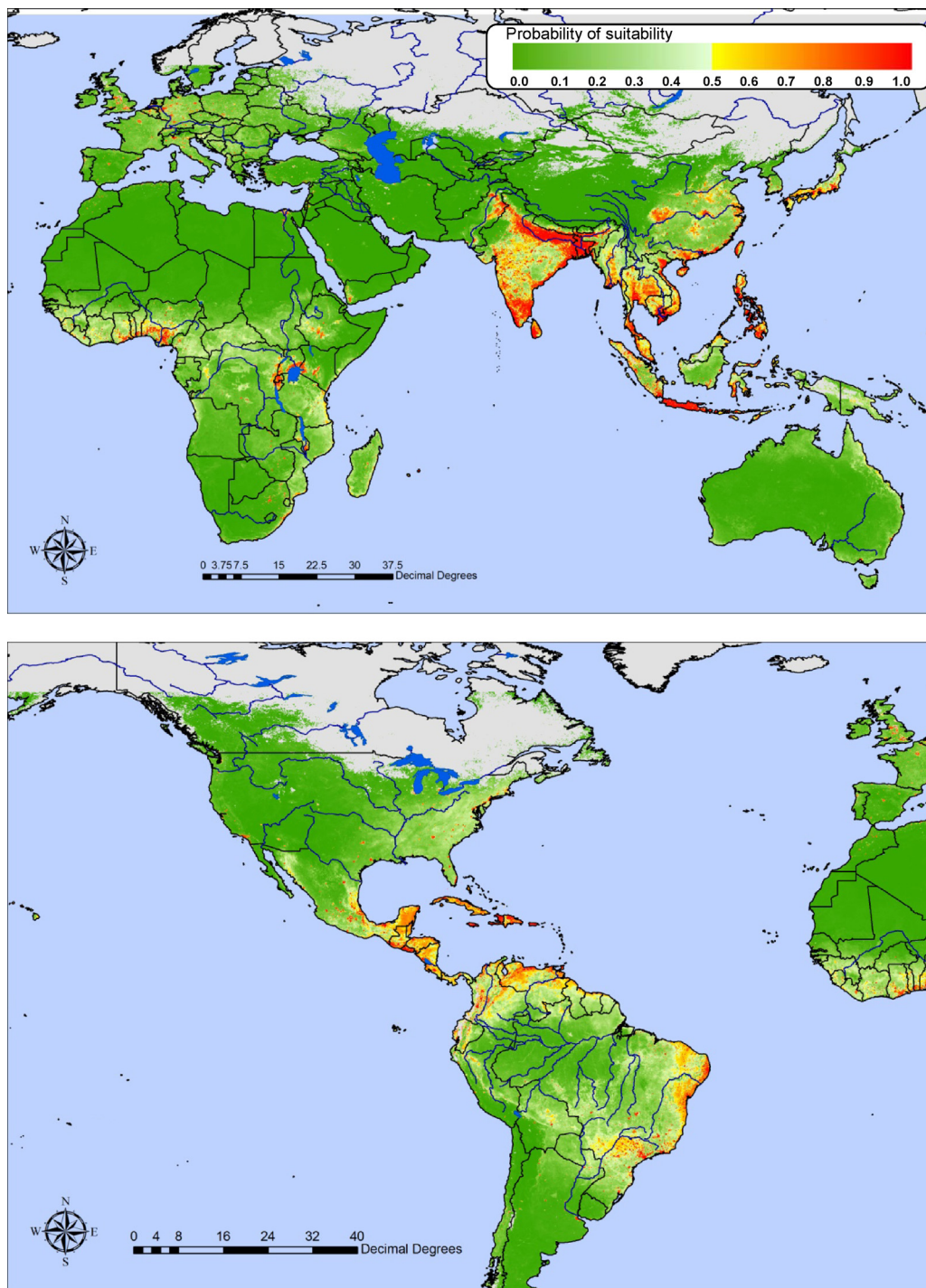


Fig. 1. Global risk map for dengue. This map was produced using the all-dengue database and included the modelled distributions of dengue's two vector species, *A. aegypti* and *A. albopictus*, as potential predictor data layers (not selected in all models). The map shows the 'probability of suitability', i.e. the similarity between each pixel and any of the presence clusters included in the model. The grey areas are so different from any of the presence or absence clusters that no predictions are made for them.

the variables selected, and their mean values. The rainbow file from this series of models is shown in Fig. 2 and the overall ranking of the top variables selected across the 100 bootstrap models is shown in Table 3. Various accuracy statistics for the different dengue model series are shown in Table 4.

The global map (Fig. 1) captures well the global distribution of dengue, although positive predictions are sparse for central South America (the Amazon basin) and in Africa, despite the far more widespread predictions in the same areas of the presence of dengue's two vector species (the pattern of dengue predictions is

similar even when the mosquito risk maps are omitted from the predictor dataset). The reason for this appears to be the absence of high population densities of humans in both areas. Table 2 shows the very great difference between the average human population density in areas with and without dengue (overall mean values of 1024.6 vs 30.5 humans per square kilometre) for the top model and Table 3 shows that across all 100 bootstrap models human population density ranked second overall (after mean daytime Land Surface Temperature) as the best predictor of dengue's presence globally. Given its importance, and frequent selection in these

Table 2
Selected variables and their mean values in the five presence and five absence clusters of the best all-dengue model.

	wj1507a0	wj1508mn	GRUMP	wj1503mn	AEALB	mmph50a2	wj1507p3	wj1514p2	wj1507p3	wj1514p2	n (sample)
C1	32.01	9.6	55.86	0.16	0.37	25.12	2.01	2.51	0.53	1.93	146
C2	16.9	-8.29	18.57	0.12	0.15	10.1	2.12	2.11	0.54	0.69	39
C3	29.21	15.02	20.31	0.13	0.51	33.94	1.85	3.55	0.44	2.59	325
C4	25.89	-5.16	31.88	0.19	0.2	9.89	1.87	2.26	0.54	1.54	53
C5	21.24	-10.87	30.09	0.14	0.04	5.43	1.93	2.93	0.54	2.17	30
ALL ABSENT	28.39	9.037	30.476	0.142	0.402	26.612	1.912	3.054	0.483	2.188	593
C6	23.12	14.07	1205.57	0.06	0.61	40.63	2.07	3.33	0.46	3.24	90
C7	30.51	7.43	1512.17	0.12	0.42	27.27	1.97	2.43	0.54	1.62	44
C8	31.92	15.61	716.46	0.1	0.56	24.61	2.49	3.24	0.49	2.61	155
C9	-7.63	11.05	3177.84	0.08	0.29	40.7	1.91	2.92	0.51	2.74	30
C10	28.75	17.03	836.39	0.08	0.69	30.66	2.25	3.42	0.42	3.37	290
ALL PRESENT	27.06	15.244	1024.591	0.086	0.603	30.842	2.248	3.265	0.456	2.999	609
ALL (abs. and pres.)	27.716	12.182	534.15	0.114	0.504	28.755	2.082	3.161	0.47	2.599	1202

Key: C1–C5: absence clusters; C6–C10: presence clusters. wj1507a0, mean daytime Land Surface Temperature, degrees Celsius; wj1508mn, minimum night-time Land Surface Temperature, degrees Celsius; GRUMP, human population density, per square km; wj1503mn, minimum Middle Infra Red, reflectance; AEALB, posterior probability prediction of the presence of *A. albopictus*; mmph50a2, bi-annual amplitude of the CMORPH rainfall estimate, mm; wj1507p3, tri-annual phase of day-time Land Surface Temperature, decimal months from 1st January; wj1514p2, bi-annual phase of the Normalised Difference Vegetation Index, decimal months; wj1508a3, tri-annual amplitude of the night-time Land Surface Temperature, degrees Celsius; wj0857p2, bi-annual phase of WORLDCLIM rainfall estimate, decimal months; n, bootstrap cluster sample size. ALL ABSENT, weighted mean values of all absence sites. ALL PRESENT, weighted mean values of all presence sites. ALL (abs. and pres.) = weighted mean values of all sites.

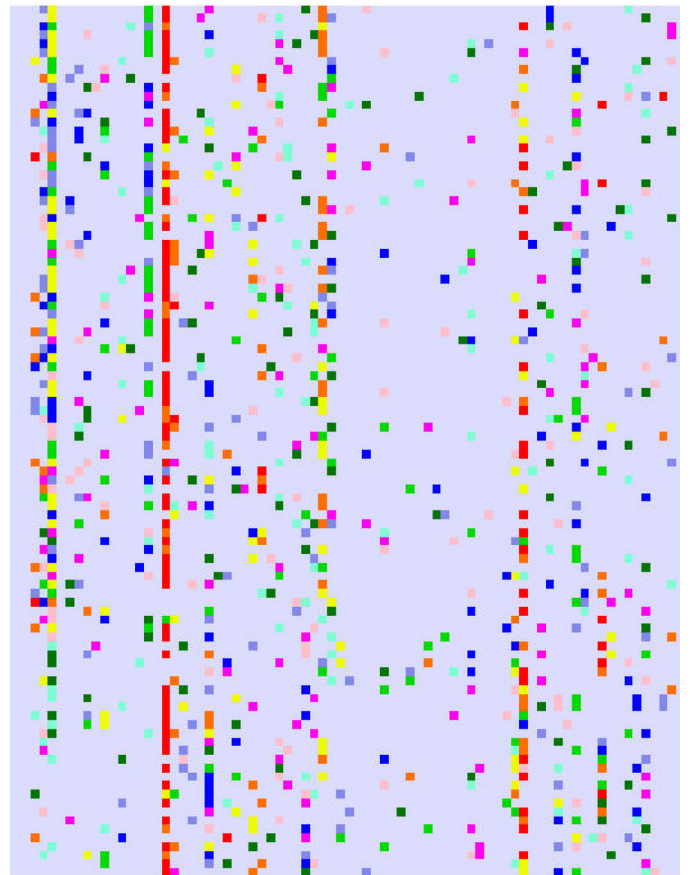


Fig. 2. Rainbow file for the all-dengue models of Fig. 1. For a description of rainbow files, see the text, Section 2.4. The first two coloured columns from the left in this image refer to *A. aegypti* and *A. albopictus*, the next to human population density (GRUMP), the next is spare and the next is digital elevation. There then follow blocks of ten columns referring to successive Fourier variables; MIR, day-time LST, night-time LST, NDVI, EVI, CMORPH and WORLDCLIM data (see text Table 1 for full names). The column with many red pixels (left of centre) is for the day-time LST mean. This variable was frequently selected as the first (red) or second (orange) variable in the bootstrap models and was overall the most important predictor variable. GRUMP (third coloured column from the left) was the second most important variable for all the models in this series, frequently selected third (yellow) or fourth (green) but was never selected first.

models, it is possible to understand the absence of positive predictions for dengue in places with low human population densities.

Table 2 shows there is a mean difference of just over one degree Celsius in the mean day-time Land Surface Temperature (selected in 80/100 models) between areas with and without dengue globally (27.06 °C vs 28.39 °C, respectively), a counter-intuitive result which is resolved by looking at the mean temperatures of individual clusters contributing to these overall mean values. When the mean temperatures of the presence and absence clusters are ranked separately there is a relatively close correspondence between the two series, with three of the five presence clusters (C7, C10 and C6) experiencing higher mean temperatures than corresponding absence clusters (C3, C4 and C5, respectively), as we would expect; one presence cluster (C8) experiencing only marginally lower mean temperatures than a corresponding absence cluster (C1); and only one presence cluster (C9) with very much lower temperatures than the remaining absence cluster (C2; -7.63 °C vs 16.9 °C, respectively). It is this latter presence cluster that brings down the global average for presence sites.

The second most important variable in Table 2, the minimum night-time Land Surface Temperature (selected in 40/100 models), shows consistently higher mean temperatures in areas of

Table 3
Mean ranking of the top 20 variables across the 100 all-dengue bootstrap models.

Position	Variable	Mean rank	<i>n</i>	Position	Variable	Mean rank	<i>n</i>
1	Day LST mean	3.26	80	11	CMORPH mean	9.57	22
2	GRUMP	6.21	84	12	Night LSTphase1	9.6	18
3	CMORPH phase 1	6.93	47	13	AEAE	9.68	18
4	Night LST minimum	7.8	40	14	Night LST variance	9.76	28
5	MIR minimum	8.61	44	15	Night LST mean	9.79	19
6	Day LSTphase3	8.78	44	16	MIR phase2	9.83	23
7	AEALB	9.17	39	17	Day LST minimum	10.04	17
8	CMORPH maximum	9.19	34	18	WDCLIMphase3	10.17	25
9	WDCLIMmean	9.42	21	19	MIR phase 1	10.25	17
10	Day LSTphase1	9.52	20	20	Night LST phase2	10.26	26

Position: 1 = top-ranked variable, 2 = second-ranked variable, etc. For naming conventions and most variable names see [Tables 1 and 2](#). AEAE, posterior probability prediction of the presence of *A. aegypti*. Mean Rank, the mean rank of each variable across all 100 bootstrap models. *n*, the number of models, out of 100, in which the variable was selected.

Table 4
Accuracy statistics for the top- and bottom-ten bootstrap models (out of 100) for each dengue model.

		kappa	%Correct	%PosCorr	%NegCorr	%False +ves	%False –ves	Sensitivity	Specificity	TSS	AUC
Dengue Fever	Top 10	0.73 (0.018)	86.4 (0.90)	80.6 (2.42)	92.2 (1.79)	3.9 (0.90)	9.7 (1.21)	0.91 (0.017)	0.82 (0.016)	0.73 (0.016)	0.94 (0.007)
	Bottom 10	0.56 (0.034)	78.2 (1.71)	82.1 (4.87)	85.4 (9.25)	14.6 (4.86)	6.9 (3.57)	0.75 (0.068)	0.84 (0.043)	0.59 (0.046)	0.88 (0.017)
Dengue Haemorrhagic Fever	Top 10	0.79 (0.028)	89.4 (1.40)	89.5 (1.81)	89.2 (2.82)	5.4 (1.40)	5.2 (0.90)	0.89 (0.024)	0.89 (0.016)	0.78 (0.028)	0.96 (0.007)
	Bottom 10	0.68 (0.027)	84.0 (1.34)	88.3 (2.81)	85.7 (3.00)	9.0 (1.99)	7.0 (0.99)	0.82 (0.028)	0.85 (0.015)	0.67 (0.027)	0.92 (0.009)
All dengue	Top 10	0.74 (0.017)	87.1 (0.83)	83.4 (1.94)	90.8 (1.87)	4.6 (0.94)	8.3 (0.98)	0.90 (0.017)	0.84 (0.015)	0.74 (0.018)	0.95 (0.005)
	Bottom 10	0.59 (0.030)	79.6 (1.52)	81.8 (4.47)	87.7 (6.91)	12.9 (3.85)	7.5 (3.23)	0.77 (0.055)	0.83 (0.038)	0.60 (0.034)	0.89 (0.016)

Key: kappa = Cohen's kappa; % Correct, % of all observations predicted correctly; %PosCorr, % correct presence predictions; %NegCorr, % correct negative predictions; %False +ves, % of incorrect predictions of presence; %False –ves, % of incorrect predictions of absence; Sensitivity, proportional correct predictions of known presences; Specificity, proportional correct predictions of known (pseudo-)absences; TSS, True Skill Statistic (=Sensitivity + Specificity – 1); AUC, Area under the curve (=ROC statistic). Standard errors are given in brackets after each mean value.

dengue presence versus absence, with averages across all clusters of 15.24°C vs 9.04°C, respectively. In our experience, minimum night-time Land Surface Temperature is frequently a key predictor variable for vectors, and hence vector-borne diseases; it is often not the maximum temperature reached during the daytime that is critical, but the minimum night-time temperature, affecting both the survival of the mosquito vectors and the extrinsic incubation period of the diseases they carry.

Table 3 shows that the top twenty key predictor variables for the all-dengue models, apart from human population density, are various temperature variables and a few rainfall-related ones, with no vegetation-related variables at all (the first vegetation-related variable appears at position 22 in the list shown partly in Table 3, with a mean rank of 10.32, and selected in only 20/100 models).

The all-dengue models chose the *A. albopictus* predictor layer in 39 of the 100 bootstrap models but the *A. aegypti* predictor layer in only 18 models, with overall mean ranks for these two species of 9.17 and 9.68, in seventh and thirteenth positions, respectively. Whilst the mean values of the *A. albopictus* predictor layer in Table 2 is as expected overall (mean posterior probabilities of 0.40 and 0.60 in areas of absence and presence, respectively, i.e. neatly straddling the 0.5 probability distinguishing predicted mosquito absence from presence), individual presence clusters (C7 and C9) have mean values below the 0.5 threshold and one absence cluster has a value just above 0.5 (C3, mean value 0.51).

The all-dengue models combine data on both dengue fever and dengue haemorrhagic fever and the models for these two components separately show interesting contrasts. Human population density and mean day-time Land Surface Temperature were the top two variables for the dengue fever risk map (selected in 74 and 52 of the 100 models, respectively) and the *A. aegypti* and *A. albopictus* predictor variables were in overall fourth and sixth positions, with mean ranks of 8.12 and 8.62 (each selected in 37 models), respectively. In contrast, mean day-time Land Surface Temperature and the annual phase of the same variable were the top two variables in the dengue haemorrhagic fever models (selected in 47 and 38 models), the two vector species were ranked in twenty-fifth and thirteenth positions (selected in 12 and 34 models) respectively and human population density was selected in only one of the 100 bootstrap models, ending in position 66 (mean rank = 10.92) out of the total of 74 predictor variables available. In the single model in which it was chosen it demonstrated a clear difference from the dengue fever models. Thus, the mean human population densities in areas of dengue absence were approximately the same in the best bootstrap model of each series (33.05 and 33.60 humans per square km, respectively) but were 1065 in areas of dengue fever (the range of mean values across the five presence clusters was 472 to 3043) and 1667 in areas of dengue haemorrhagic fever (the range of mean values across the five presence clusters was 927 to 6196). Dengue haemorrhagic fever is

particularly common in SE Asia, an area with high human population densities, but has also been recorded from the coastal areas of Central and South America where human population densities are lower.

Thus it appears that dengue fever is characterised by high human population densities more or less wherever it occurs, whilst dengue haemorrhagic fever is associated with extremely high population densities only in relatively few parts of its range, not in others. The global predictions of dengue haemorrhagic fever high risk areas included much larger parts of the coastal and moist savannah parts of Africa North of the equator than did the dengue fever, or all-dengue, models. Africa tends not to have the extremely high human population densities typical of SE Asia or India, so that these relatively widespread positive predictions for haemorrhagic fever in Africa are due to variables other than human population density, mostly thermal and rainfall ones. In contrast, the consistent importance of high human population densities for dengue fever explains why Africa is not predicted to have many high risk regions for this form of dengue; Africa does not at present have the necessary high human population densities.

The rainbow file for the all-dengue maps (Fig. 2) shows that mean daytime Land Surface Temperature, top in the best bootstrap model (Table 2) and with the highest overall rank (Table 3), is often selected as the first or second most important variable in the individual bootstrap models. The rainbow file also shows that human population density is frequently selected in the models, but not at such a high position.

The kappa model accuracy statistics in Table 4 are good to excellent for both the top- and bottom- ten models of each disease (range 0.56–0.79). The mean AUC statistic is never less than 0.88 (range 0.88–0.96) (the fact that the AUC was calculated in the same way for each model removes some of the acknowledged problems with this metric; Lobo et al., 2008). These are encouraging results bearing in mind that the global distributions of the three diseases were captured by either only three (dengue haemorrhagic fever) or only five (dengue fever and all-dengue) clusters in each of the presence and absence categories and using a total of no more than ten predictor variables for any one bootstrap model.

Table 5 shows the importance of individual variables for all-dengue, established by the method of combinations (Burnham and Anderson, 2002). Human population density is the clear winner. The mean scaled AICc for this variable is considerably less than that for the following variable, the bi-annual phase of the WORDCLIM rainfall variable. There then follows a group of three variables, all related to the variability or timing of rainfall. A mixture of mostly rainfall and temperature variables completes the table, with only one vegetation index variable (the minimum Enhanced Vegetation Index) in the top 20.

The combination method thus supports the critical importance of human population density, rainfall and temperature variables (with a greater emphasis on rainfall rather than temperature

Table 5
Importance of individual variables in describing the global distribution of all-dengue, determined by the method of combinations (see text, Section 2.5).

Position	Variable	Scaled AICc	Position	Variable	Scaled AICc
1	GRUMP	1.231	11	WDCLIM amp1	1.706
2	WDCLIM phase2	1.633	12	Day LST phase3	1.712
3	WDCLIMamp3	1.654	13	CMORPH amp2	1.714
4	WDCLIM phase3	1.654	14	CMORPH variance	1.715
5	CMORPH phase1	1.654	15	Night LSTmean	1.716
6	CMORPH maximum	1.671	16	Day LST minimum	1.721
7	NightLSTphase2	1.673	17	WDCLIM amp2	1.723
8	CMORPH phase3	1.687	18	CMORPH mean	1.728
9	EVI minimum	1.697	19	Day LST maximum	1.735
10	WDCLIM mean	1.700	20	WDCLIM maximum	1.737

Key: Scaled AICc, the scaled corrected Akaike Information Criterion. For variable names, see Tables 1–3.

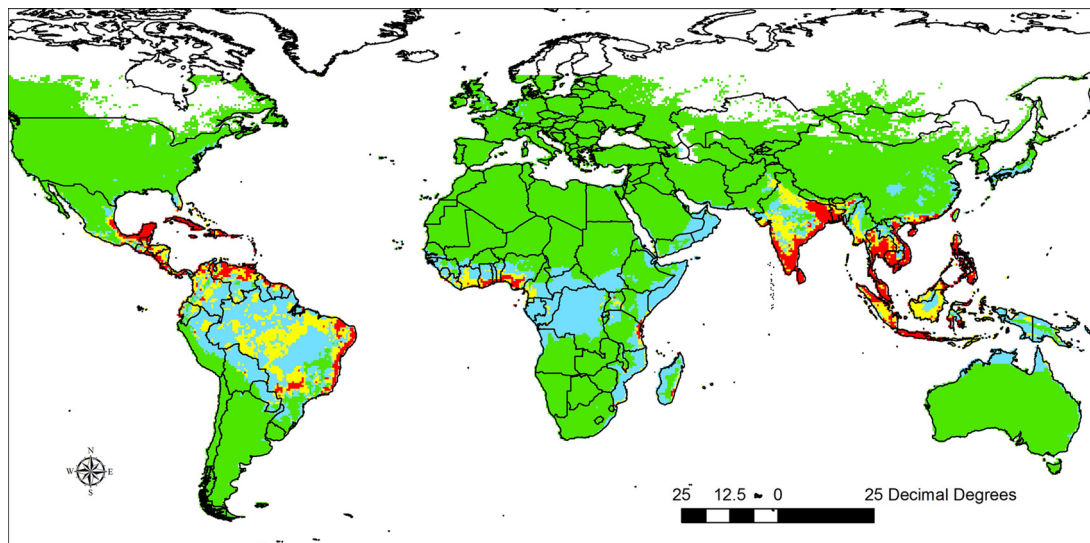


Fig. 3. Comparison between three published dengue risk maps (see text Section 2.6 for details). On this map green indicates areas where all maps predict dengue absence and red indicates where all maps predict dengue presence. Pale blue; one map predicts presence and the other two absence. Yellow; two maps predict presence and one predicts absence.

variables; contrast Tables 3 and 5), and the much lesser importance of vegetation (which is often a proxy for humidity).

Fig. 3 shows the consensus between three published global dengue predictions, and Table 6 shows the calculated Fleiss kappa statistics for the three-way comparisons, and Cohen's kappa statistic for various two-way comparisons. Fleiss's kappa (0.377) indicates only fair agreement between the three global dengue predictions. Cohen's kappa indicates poor agreement between the present map and that of Hales et al. (2002) but good agreement with that of Bhatt et al. (2013), and also good agreement between these latter two maps. The best regional agreement between the maps is in India and SE Asia (moderate agreement between all three maps and good agreement for all pairwise comparisons), the next best in the Americas (fair agreement between all three maps but poor to good agreement in pairwise comparisons) and the worst in Africa (only slight agreement between all three maps and poor agreement in all pairwise comparisons) (Landis and Koch, 1977).

Fig. 3 shows that the main differences between the maps are in the Amazon and Congo River basins, and in central India (where only one or two of the three maps predict the presence of dengue), and around the fringes of dengue's predicted global distribution (where, in any one place, only one of the maps predicts the presence of dengue). Some of the differences between these maps are attributable to the fact that human population density was not used at all as a predictor in the map of Hales et al. (2002) nor directly in the map of Bhatt et al. (2013) which, instead, used a categorical map of rural, peri-urban and urban areas. The models presented here suggest that these areas are indeed climatically suitable, but lack the high human densities usually associated with dengue. This association is probably causal rather than correlative; in high density conditions human beings create many

breeding sites suitable for *A. aegypti*, a species closely associated with humans globally.

The European part of the global dengue risk map is shown in Fig. 4. Most of Europe is predicted unsuitable for dengue but many large cities or highly populated areas are highlighted as suitable. High human population density (an important variable in the all-dengue models) is not sufficient on its own to make any area suitable for this disease and other key variables, such as day- and night-time Land Surface Temperatures (Table 3), are also playing a role in these predictions. A further series of models was run without the two vector species being included in the predictor dataset. In this series, the same variables occupied the top three positions, mean daytime Land Surface Temperature (selected in 94 bootstrap models), human population density (selected in 73 models) and the annual phase of the CMORPH rainfall variable (selected in 53 models). The average of the 100 models again picked out major European cities as at some risk of dengue.

Whilst most of Europe is presently unsuitable for dengue it is important to establish where dengue might first appear, and where it might spread. Fig. 5 shows the Mahalanobis distance image from the all-dengue model, for the same area as shown in Fig. 4 and, for comparison, Fig. 6 shows part of the same distance image for the southern United States, where dengue occurred in the relatively recent past. Fig. 6 also shows the sites where dengue occurs today (these points form part of the all-dengue database) and it is possible to see that they fall within the pale blue or beige areas of the image, indicating low Mahalanobis distances. Clearly areas in the southern USA have environmental conditions very similar to those where dengue still occurs. High living standards have banished dengue from the United States, but not the environmental conditions where

Table 6
Consensus between 3 global dengue risk maps.

	Fleiss	Map 1 vs Map2	Map1 vs Map3	Map2 vs Map3
Global	0.377	0.295	0.522	0.403
Americas	0.325	0.230	0.399	0.428
Africa	0.095	0.082	0.330	0.139
SE Asia	0.566	0.531	0.623	0.557

Key: Fleiss, the Fleiss kappa comparing all three maps together; Map1 vs Map2, etc., Cohen's kappa index for pairwise comparisons. Map1, this publication; Map 2, from Hales et al. (2002); Map3, from Bhatt et al. (2013).

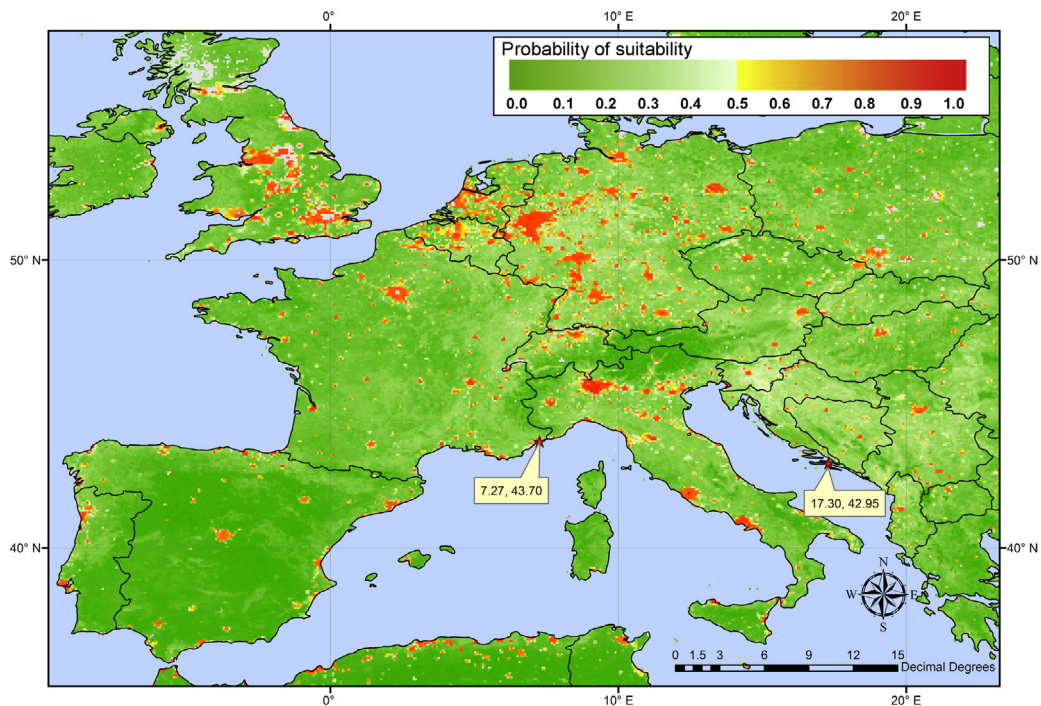


Fig. 4. Detail of Fig. 1 for Europe. Most of Europe is predicted unsuitable for dengue, but the red areas here, major towns and cities, indicate suitability for dengue. These predictions are mainly driven by both high human population densities in cities and thermal conditions that coincide with other places globally where dengue occurs. The two red asterisks with callouts indicate where local transmission of dengue has recently been reported in Europe.

dengue thrives. Fig. 5 shows that there are areas in Europe which are quite similar (in Mahalanobis distance values) to those of the southern USA, especially in northern Italy, parts of Austria, Slovenia and Croatia, and West of the Alps in France; other areas of concern include Bosnia and Herzegovina, Serbia and Montenegro,

Albania, Greece and parts of both Switzerland and Germany. Particularly worrying are those parts of these areas that have recently been invaded by *A. albopictus* (*A. aegypti* does not occur in mainland Europe), especially northern Italy, and these areas should certainly be monitored carefully in the future.

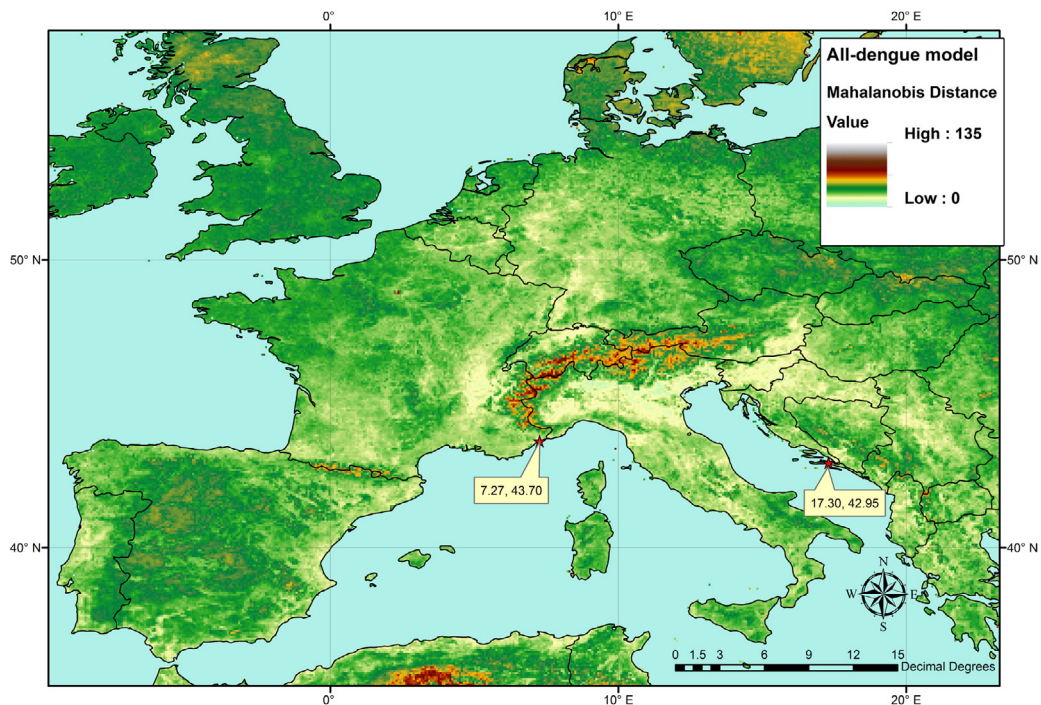


Fig. 5. Mahalanobis Distance image for Europe. This is an image of the distance in environmental space between each pixel in this image and any area globally where dengue has been recorded in the past (see text Section 2.7 for an explanation). In this map the very pale blue areas, and the lighter beige areas are quite similar to areas where dengue already occurs (i.e. low Mahalanobis Distances), and so should be regarded as the most likely places where dengue might first return to Europe. The two red asterisks, where local transmission of dengue has recently been reported in Europe, are within or very close to such areas.

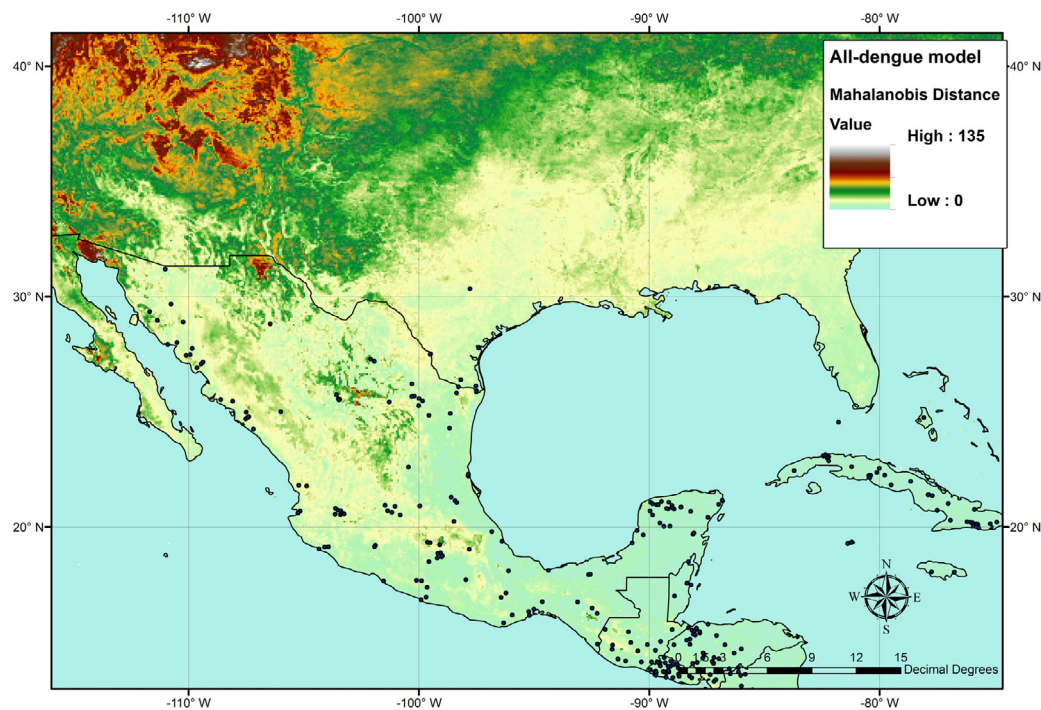


Fig. 6. Mahalanobis Distance image for North America. This is part of the same global image as shown in Fig. 5. The dots on this image show positive dengue records from the all-dengue database. Notice that many parts of southern USA, where dengue occurred historically, show Mahalanobis Distances similar to those where dengue has been reported more recently. These areas are similar environmentally to the similarly coloured areas in Fig. 5.

4. Discussion and conclusions

Many species' distribution modelling approaches involve some sort of data mining to match the pattern of points in a database to sets of environmental and other predictors. It is a truism that any pattern can be matched as long as sufficient variables, thresholds and break-points are allowed in the models. These models therefore rely heavily on representative databases; every part of a species' or disease's distribution should be represented in the database, preferably without any regional bias. Unfortunately, virtually no database of this sort is without bias and so the question becomes one of whether we are describing the distribution of a set of points (the training set) or of a species or disease. Arguably, the more flexible is the modelling approach, the more likely it is that the distribution of the points rather than the species will be described.

Biologically based predictions of species' distributions start out with a set of assumptions about the habitat (temperature, humidity, etc.) conditions that the species requires, and make predictions based on these. Such models require absolutely no training set data (i.e. point records of occurrence) of any sort, and a mark of how good they are is to see how well their predictions coincide with what is known of the distribution of the modelled species. Unfortunately very little is known about the real environmental constraints of many species or diseases and, even when something is known, critical thresholds and developmental and other rates are often derived from laboratory studies, which may be both unreliable and incomplete. Missing out a key variable, because laboratory or other studies have not yet been made of it, is likely to produce a poor distribution prediction, as is the omission of important biotic constraints such as hosts, parasites or predators (Soberon and Peterson, 2005).

The middle way between a purely data mining and a 'purist' biological approach is one that assumes certain rather simple constraints on a species' performance, for example that a species' response to a gradually changing variable, such as temperature gradually rising from a low value, is first positive and later negative,

as lethal temperatures are reached. Without specifying in advance which are the most important variables, this assumed response curve can be used to inform the modelling approach. Clearly, simple linear logistic equations, which were enormously important in the early days of species' distribution modelling, incorporate part of the presumed overall response, but not all of it (Rogers, 2006). Normal and multi-variate normal curves are tractable statistically and appear to give a good first order description of species' responses to environmental conditions. They are at the heart of the non-linear discriminant analysis approach used for the models presented here. By electing to choose only a limited set of environmental predictors, and no threshold or breakpoints, the resulting models should be more open to biological interpretation than the models produced by more complicated, pattern matching approaches.

The present predictions for dengue globally differ from those produced by one of us (DJR) previously (Rogers et al., 2006) in predicting less widespread areas of dengue risk in both Africa and South America. The main difference between the present and previous models was the use in the present models of human population density as a predictor variable. The previous study highlighted a coincidence between areas of high human population density and areas predicted to be at high risk of dengue (with yellow fever providing the exact contrast, of usually being associated only with low human population densities), but did not use human density as a predictor in the models (Fig. 10 in Rogers et al., 2006). The strong association of dengue fever with high human population densities means that when such densities are used as predictors in dengue models, the predicted high risk areas coincide much more with human population distribution, and the relative importance of environmental predictors is diminished, but not eliminated entirely: thermal and rainfall conditions still remain crucial. This is especially noticeable in Europe where cities appear to be at some degree of risk, but rural areas are not, at the present time.

Of equal interest is the fact that both the vector species' predictor layers were relatively rarely chosen in the dengue haemorrhagic fever models, and human population density appeared in only one

of the 100 bootstrap models in this series. In this single bootstrap model it conformed to expectations; human densities are very much higher in areas of DHF risk, even compared with areas of dengue fever risk. But the infrequent choice of both vectors and humans as predictors of DHF (as opposed to predictors of dengue fever), suggests that some other conditions define DHF areas, and that DHF currently occurs in areas of a wide and variable range of human population densities. Severe dengue (DHF) tends to occur only in patients who previously suffered from dengue fever caused by a heterologous strain of dengue. For an area to have DHF it must therefore have in circulation at least two of the four distinct serotypes of dengue, historically a relatively rare event. For example, before 1970 only nine countries had experienced outbreaks of DHF; less than thirty years later this number had risen to 102 (Dorji et al., 2009). Such a rapid and recent spread suggests that DHF has not yet settled to its final pattern of distribution, making modelling of it difficult at the present time.

The relatively poor overall agreement between three published maps of predicted dengue risk is a cause for concern, because there must consequently be great uncertainty arising out of predictions drawn from any one of them. All three maps had respectable accuracy statistics using the metrics common to all of them (this paper, Table 4; a best model overall accuracy of 92% with a sensitivity of 0.85 and specificity of 0.93 for Hales et al. (2002); and an ensemble overall accuracy of 75% with a sensitivity of 0.69 and specificity of 0.81 for Bhatt et al. (2013)), but their consensus at the global level was only 'fair' and, even regionally, did not exceed 'moderate' on the scale of Landis and Koch (1977) (Table 6). No pair-wise comparison of any two of the maps exceeded a 'good' agreement (that is, a Cohen's kappa of >0.75) either globally or regionally (Table 6). How can maps that are individually good have such poor agreement between them? Part of the reason for this might be that Boosted Regression Trees (BRTs, the method used in Bhatt et al., 2013) are good at describing exceptional or unusual data points, whilst the methods used in the present paper, and by Hales et al. (2002), concentrate on trying to find a simple set of (multi-variate normal or logistic regression) rules to describe the data; each method is likely to miss certain points, but they will be different points.

Mahalanobis distance images are recommended as a way of showing the potential environmental suitability of sites for vectors or diseases, and can also be used in monitoring programmes. Whilst the present global dengue risk map is informed by training set points where the vector was almost always *A. aegypti*, the most likely vector for dengue in Europe is *A. albopictus*. Historically dengue fever (never DHF) outbreaks have only infrequently been associated with this vector (Lambrechts et al., 2010), but *A. albopictus* is a flexible and adaptable species, as its involvement in the chikungunya outbreak in Europe clearly shows (Pialoux et al., 2007). Particularly worrying therefore are the areas in the Mahalanobis distance image where low values (indicating environmental suitability for dengue) coincide with areas where *A. albopictus* is already well established, for example in northern Italy. The two places where autochthonous transmission has recently occurred in mainland Europe (i.e. France and Croatia, indicated in Figs. 4 and 5) are close to, or within, low Mahalanobis Distance regions of Fig. 5, but it should be stressed that this distance image is based on a global dengue database where most transmission is by *A. aegypti*, not by *A. albopictus* that is almost certainly the vector in these two sites. It is impossible to know at this stage what are the likely errors of using a risk map based on data mostly or exclusively involving one vector species to make predictions for the same disease transmitted by a different vector, and in different places. Undoubtedly the former may be taken as a guide to the latter, until further information becomes available but, given that most parameters in vector-borne disease models are related

in some way to vector (rather than host) biology, such extrapolation should be treated cautiously.

Perhaps the best way forward is to maintain and expand the current databases of both vectors and diseases (e.g. Brady et al., 2012) and to generate both risk maps and Mahalanobis distance images of vectors and diseases to see where high risk areas for the vectors coincide with high risk areas for the disease. These sites should then be monitored periodically for signs of the vectors and diseases. One lesson of history is that it is more or less impossible to predict where a new vector or disease will arrive in a previously uninfested region. But, having arrived, it should be possible to predict whether or not that vector or disease will survive there, and the direction in which (and possibly the speed at which) it will spread. Vectors and diseases move over both physical and environmental landscapes. Failure to appreciate this will render us unable fully to understand, and therefore predict, the future of vector-borne and other environmentally sensitive diseases.

Conflicts of interest statement

There are no conflicts of interest arising from this work.

Acknowledgements

Much of this work was performed under a direct service contract, OJ/12/12/2008-PROC/2008/044, issued by the European Centre for Disease Prevention and Control (ECDC). We are grateful to Hervé Zeller, Wim Van Bortel and Bertrand Sudre of the ECDC for support and discussions during and after the analyses described here, and to this special issue editor, Mercedes Pascual, for comments on the manuscript. Becky Hay kindly organised and undertook expanding a pre-existing dengue database for this study, and established new ones for both vector species, to a state where they could be used in the present exercise, and Will Temperley provided vital database management services. The bank of temporal Fourier imagery was produced and maintained by David Benz.

Finally, the authors are grateful to both Simon Hales and Oliver Brady for the supply in GIS format of the global dengue maps that appeared in the original publications (Hales et al., 2002; Bhatt et al., 2013).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.actatropica.2013.08.008>.

References

- Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *J. Appl. Ecol.* 43, 1223–1232.
- Balk, D.L., Deichmann, U., Yetman, G., Pozzi, F., Hay, S.I., Nelson, A., 2006. Determining global population distribution: methods, applications and data. *Adv. Parasitol.* 62, 119–156.
- Barve, N., Barve, V., Jimenez-Valverde, A., Lira-Noriega, A., Maher, S.P., Peterson, A.T., Soberon, J., Villalobos, F., 2011. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecol. Model.* 222, 1810–1819.
- Benedict, M.Q., Levine, R.S., Hawley, W.A., Lounibos, L.P., 2007. Spread of the tiger: global risk of invasion by the mosquito *Aedes albopictus*. *Vector Borne Zoonotic Dis.* 7, 76–85.
- Bhatt, S., Gething, P.W., Brady, O.J., Messina, J.P., Farlow, A.W., Moyes, C.L., Drake, J.M., Brownstein, J.S., Hoen, A.G., Sankoh, O., Myers, M.F., George, D.B., Jaenisch, T., Wint, G.R., Simmons, C.P., Scott, T.W., Farrar, J.J., Hay, S.I., 2013. The global distribution and burden of dengue. *Nature* 496, 504–507.
- Brady, O.J., Gething, P.W., Bhatt, S., Messina, J.P., Brownstein, J.S., Hoen, A.G., Moyes, C.L., Farlow, A.W., Scott, T.W., Hay, S.I., 2012. Refining the global spatial limits of dengue virus transmission by evidence-based consensus. *PLoS Negl. Trop. Dis.* 6 (8), e1760, <http://dx.doi.org/10.1371/journal.pntd.0001760>.
- Burnham, K.P., Anderson, D.R., 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. Springer, New York.

- Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. *Rem. Sens. Environ.* 37, 35–46.
- Degallier, N., Favier, C., Menkes, C., Lengaigne, M., Ramalho, W.M., Souza, R., Servain, J., Boulanger, J.-P., 2010. Toward an early warning system for dengue prevention: modeling climate impact on dengue transmission. *Clim. Change* 98, 581–592.
- Dorji, T., Yoon, I.K., Holmes, E.C., Wangchuk, S., Tobgay, T., Nisalak, A., Chinnawirotpisan, P., Sangkachantaranon, K., Gibbons, R.V., Jarman, R.G., 2009. Diversity and origin of dengue virus serotypes 1, 2, and 3, Bhutan. *Emerg. Infect. Dis.* 15, 1630–1632.
- Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Townsend Peterson, A., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberon, J., Williams, S., Wisz, M.S., Zimmermann, E., 2006. Novel methods improve predictions of species' distributions from occurrence data. *Ecography* 29, 129–151.
- Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. *J. Anim. Ecol.* 77, 802–813.
- Fichet-Calvet, E., Rogers, D.J., 2009. Risk maps of Lassa fever in West Africa. *PLoS Negl. Trop. Dis.* 3 (3), e388, <http://dx.doi.org/10.1371/journal.pntd.0000388>.
- Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76, 378–382.
- Gjenero-Margan, I., Aleraj, B., Krajcar, D., Lesnikar, V., Klobucar, A., Pem-Novosel, I., Kurecic-Filipovic, S., Komparak, S., Martic, R., Duricic, S., Betica-Radic, L., Okmadzic, J., Vilibic-Cavlek, T., Babic-Erceg, A., Turkovic, B., Avsic-Zupanc, T., Radic, I., Ljubic, M., Sarac, K., Benic, N., Mlinaric-Galinovic, G., 2011. Autochthonous dengue fever in Croatia, August–September 2010. *Eur. Surveill.* 16 (9), pii=19805. Available online: <http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19805>
- Green, P.E., 1978. *Analyzing Multivariate Data*. The Dryden Press, Hinsdale, Illinois.
- Gubler, D., 2006. Dengue/dengue haemorrhagic fever: history and current status. In: *Novartis Foundation Symposium 2006*, pp. 3–16, discussion 16–22, 71–13, 251–253.
- Hales, S., de Wet, N., Maindonald, J., Woodward, A., 2002. Potential effect of population and climate changes on global distribution of dengue fever: an empirical model. *Lancet* 360, 830–834.
- Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* 25, 1965–1978.
- Jetten, T.H., Focks, D.A., 1997. Potential changes in the distribution of dengue transmission under climate warming. *Am. J. Trop. Med. Hyg.* 57, 285–297.
- Joyce, R.J., Janowiak, J.E., Arkin, P.A., Xie, P., 2004. CMORPH: a method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *J. Hydrometeorol.* 5, 487–503.
- Krzanowski, W.J., Marriott, F.H.C., 1995. *Multivariate Analysis. Part 2. Classification, Covariance Structures and Repeated Measurements*. Arnold, London.
- La Roche, G., Souares, Y., Armengaud, A., Peloux-Petiot, F., Delaunay, P., Despres, P., Lenglet, A., Jourdain, F., Leparac-Goffart, I., Charlet, F., Ollier, L., Mantey, K., Mollet, T., Fournier, J.P., Torrents, R., Leitmeyer, K., Hilairet, P., Zeller, H., Van Bortel, W., Dejour-Salamanca, D., Grandadam, M., Gastellu-Etchegorry, M., 2010. First two autochthonous dengue virus infections in metropolitan France, September 2010. *Eur. Surveill.* 15 (39), pii=19676. Available online: <http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19676>
- Lambrechts, L., Scott, T.W., Gubler, D.J., 2010. Consequences of the expanding global distribution of *Aedes albopictus* for dengue virus transmission. *PLoS Negl. Trop. Dis.* 4 (5), e646, <http://dx.doi.org/10.1371/journal.pntd.0000646>.
- Landis, J.R., Koch, G.C., 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174.
- Lobo, J.M., Jimenez-Valverde, A., Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecol. Biogeogr.* 17, 145–151.
- Macarthur, R.H., 1972. *Geographical Ecology*. Harper & Row, New York.
- McPherson, J.M., Jetz, W., Rogers, D.J., 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *J. Appl. Ecol.* 41, 811–823.
- Patz, J.A., Martens, P., Focks, D.A., Jetten, T.H., 1998. Dengue fever epidemic potential as projected by general circulation models of global climate change. *Environ. Health Perspect.* 106, 147–153.
- Pfeiffer, D.U., Robinson, T.P., Stevenson, M., Stevens, K.B., Rogers, D.J., Clements, A.C.A., 2008. *Spatial Analysis in Epidemiology*. Oxford University Press, Oxford.
- Phillips, S.J., Dudik, M., Schapire, R.E., 2004. A maximum entropy approach to species distribution modeling. In: *21st International Conference on Machine Learning*, Banff, Canada, 8 pp.
- Pialoux, G., Gauzere, B.A., Jaureguiberry, S., Strobel, M., 2007. Chikungunya, an epidemic arbovirovirus. *Lancet Infect. Dis.* 7, 319–327.
- Randolph, S.E., Rogers, D.J., 2010. The arrival, establishment and spread of exotic diseases: patterns and predictions. *Nat. Rev. Microbiol.* 8, 361–371.
- Reiter, P., 2010. Yellow fever and dengue: a threat to Europe? *Eur. Surveill.* 15 (10), pii=19509. Available online: <http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19509>
- Rogers, D.J., 2000. Satellites, space, time and the African trypanosomiasis. *Adv. Parasitol.* 47, 129–171.
- Rogers, D.J., 2006. Models for vectors and vector-borne diseases. *Adv. Parasitol.* 62, 1–35.
- Rogers, D.J., Sedda, L., 2012. Statistical models for spatially explicit data. *Parasitology* 139, 1852–1869.
- Rogers, D.J., Williams, B.G., 1993. Monitoring trypanosomiasis in space and time. *Parasitology* 106, S77–S92.
- Rogers, D.J., Williams, B.G., 1994. Tsetse distribution in Africa: seeing the wood and the trees. In: *Edwards, P.J., May, R.M., Webb, N.R. (Eds.), Large-Scale Ecology and Conservation Biology. 35th Symposium of the British Ecological Society with the Society for Conservation Biology (1993)*. Blackwell Scientific Publications/University of Southampton, UK, pp. 249–273.
- Rogers, D.J., Wilson, A.J., Hay, S.I., Graham, A.J., 2006. The global distribution of yellow fever and dengue. *Adv. Parasitol.* 62, 181–220.
- Scharlemann, J.P.W., Benz, D., Hay, S.I., Purse, B.V., Tatem, A.J., Wint, G.R.W., Rogers, D.J., 2008. Global data for ecology and epidemiology: a novel algorithm for temporal Fourier processing MODIS data. *PLoS ONE* 3 (1), e1408, <http://dx.doi.org/10.1371/journal.pone.0001408>.
- Simmons, C.P., Farrar, J.J., Nguyen, V., Wills, B., 2012. Dengue. *N. Engl. J. Med.* 366, 1423–1432.
- Soberon, J., Peterson, A.T., 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodivers. Inform.* 2, 1–10.
- Stevens, K.B., Pfeiffer, D.U., 2011. Spatial modelling of disease using data- and knowledge-driven approaches. *Spat. Spatiotemporal Epidemiol.* 2, 125–133.
- Stockwell, D., Peters, D., 1999. The GARP modelling system: problems and solutions to automated spatial prediction. *Int. J. Geogr. Inf. Sci.* 13, 143–158.
- Tatem, A.J., Hay, S.I., Rogers, D.J., 2006a. Global traffic and disease vector dispersal. *Proc. Natl. Acad. Sci. U.S.A.* 103, 6242–6247.
- Tatem, A.J., Rogers, D.J., Hay, S.I., 2006b. Global transport networks and infectious disease spread. *Adv. Parasitol.* 62, 293–343.