doi:10.1038/nature12060

Table of Contents

A: Assembly of the dengue database....3

A.1: Overview....3
A.2: Peer-reviewed literature search....3
A2.1: Data collection....3
A.2.2: Assigning geo-positions to data from the peer-reviewed literature....4
A.3: Collation of online informal data sources....5
A3.1: Data collection....5
A.3.2: Assigning geo-positions to data from online informal data sources....6
A.4: Automatic validation and quality control....7
A.5 Temporal standardisation....8
A.6 Data summary....8

B: Explanatory covariates....21

B.1: Overview....21
B.2: Covariate selection....22
B.2.1: Climatic and environmental covariates....22
B.2.2: Socio-economic covariates....23
B.3: Covariate sources....25
B.3.1: WorldClim database - precipitation....25
B.3.2: Temperature Suitability....26
B.3.3: Advanced Very High Resolution Radiometer - NDVI....28
B.3.4: Global Rural Urban Mapping Project....29
B.3.5: Urban Accessibility....30
B.3.6: Relative Poverty....30
B.4: Raster Standardisation....32
B.5 Covariate Extraction....33
B.6: Multicollinearity....33

C: Predicting probability of dengue transmission using Boosted Regression Trees 34

C.1: Overview....34
C.2: Boosted Regression Trees....37
C.2.1: Regression trees and boosting: a conceptual description....37

C.2.2: BRT parameter selection....38
C.2.3: Summarising the BRT model....38
C.2.4: Evaluating the BRT model predictive performance....39
C.3: Pseudo-data generation....41
C.3.1: Geographical extent....42
C.3.2: Ratio of pseudo-absences to presences....43
C.3.3: Contamination bias....43
C.3.4: Sampling Bias....44
C.3.5: Pseudo-data generation process....44
C.4: Ensemble analysis....45
C.5: Overview of Map Generation....46

C.6: Output maps and partial dependence plots **46**

D: Global burden and population-at-risk estimation 50

D.1: Overview **50**
D.2: Assembly of cohort studies **55**
D.2.1: Existing incidence data **55**
D.2.2: Inclusion criteria **56**
D.2.3: Summary **57**
D.3: Relationship between incidence and probability of occurrence **58**
D.3.1: Data model **58**
D.3.2: Process model **59**
D.3.3: Parameter model **60**
D.3.4: Posterior inference **60**
D.4 Overview of map generation and burden estimates **61**

E: Reconciling cartographic and surveillance-based burden estimates 63

E.1 Overview **63**
E.2 Surveillance-based burden data sources **63**
E.3 Country-level burden estimates **65**
E.4 Comparing cartographic and surveillance-based burden estimates **74**

F: References 81

A: Assembly of the dengue database

A.1: Overview

The dengue database comprises occurrence data linked to point or polygon locations, derived from (i) the peer-reviewed literature and case reports and (ii) HealthMap data[1]. Both data sources are described in full here. To collate the peer-reviewed database, literature searches were undertaken using major search engines and the resulting articles were manually reviewed. For the HealthMap data, online informal data sources were monitored, including online news aggregators, eyewitness reports, expert-curated discussions and validated official reports. All entries from both data sources were manually checked by the authors and then underwent a series of quality-control procedures described below to ensure correct geo-positioning. In total, 8,309 geo-positioned data points were incorporated into the modelling work described in this paper.

A.2: Peer-reviewed literature search

A2.1: Data collection

PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) 1920 to 2009 was searched using the term "dengue". The MESH term technology used in the PubMed citation archive ensured all pseudonyms were automatically included (http://www.nlm.nih.gov/mesh/2008/MBrowser.html) in the searches. The same process was repeated for ISI Web of Science (http://wok.mimas.ac.uk) and PROMED (http://www.promedmail.org). The searches were last updated on 8th February 2012. No language restrictions were placed on these searches; however, only those citations with a full title and abstract were retrieved. This resulted in a collection of 5,876 references, of which 2,883 unique articles were identified as potentially containing useable location data. The full

texts were obtained for 2,838 of these (98.4%) and the information from 1,655 articles was ultimately included in our database. The references are listed in the Supplementary Information of Brady *et al* 2012[2].

In-house language skills allowed processing of all English, French, Portuguese and Spanish articles. We were unable to extract information from a small number of Turkish, Polish, Hebrew, Italian, German and Chinese articles.

Clinical or laboratory confirmation of dengue virus transmission found within these articles was recorded as a dengue occurrence data point. Reports of autochthonous (locally transmitted) cases or outbreaks were entered as an occurrence within the country in which transmission occurred. If imported cases were reported with information on the site of contagion, they were geo-positioned to the country of contagion. If imported cases were reported with no information about the site of contagion, they were not entered into the database. If an imported case led to an outbreak (i.e. local transmission) within the recipient country and location information was available for the site of initial contagion and the site of the outbreak, this was recorded as two occurrences: one in the country of contagion and one in the country where the outbreak occurred.

A.2.2: Assigning geo-positions to data from the peer-reviewed literature

All available location information was extracted from each peer-reviewed article and PROMED case report. The site name was used together with all contextual information provided about the site position to determine its latitudinal and longitudinal coordinates using Google Maps (https://www.maps.google.co.uk/). Place names are often duplicated within a country, so the contextual information was used to ensure the right site was selected. When

the site name was not found, the contextual information was used to scan sites in the approximate area to check for names that had been transliterated in Google Maps in a different way to the published article (e.g. Imichli and Imishly).

If the study site could be geo-positioned to a specific place, it was recorded as a point location. If the study site could only be identified at an administrative area level (e.g. province or district, etc.), it was recorded as a polygon along with an identifier of its administrative unit. All formal occurrence records underwent temporal standardisation (see A.5) to ensure consistent occurrence point definitions before undergoing the quality-control process (A.4).

A.3: Collation of online informal data sources

A3.1: Data collection

Informal online data sources were collated automatically by the web-based system HealthMap as described elsewhere[1]. Briefly, HealthMap is an online infectious disease outbreak-monitoring system that captures data from a range of electronic sources in nine different languages. The system performs hourly scans of online news aggregators, listservs, electronic disease surveillance networks and public health outbreak report feeds. It captures four fields: headline (the headline, title or subject line), date (publication date), description (a brief summary), and info text (the main content of the article or report). The info text is passed to HealthMap's classification engine, which parses out one or more disease names and outbreak locations using dictionaries of disease and location patterns. The system then uses a separate algorithm to assign relevance scores that classify alerts as (i) breaking (information about a new outbreak or new information about an on-going outbreak), (ii) context (content about research, policy or background on a particular disease), (iii) warning (articles that warn

about the potential for an outbreak), (iv) not disease-related (articles that are captured by the system because they contain words that match disease names in the dictionary but are not in fact about an infectious disease) or (v) old news (an article that mentions a historical outbreak). Finally, HealthMap handles duplicates by aggregating together highly similar alerts such as those released by a news wire service and published in multiple periodicals. The requirements for including a dengue occurrence record from the HealthMap data set in our database were that the article or report contained the keywords "dengue", "dengue fever", "dengu" or "dhf" and was classified by the system as "breaking".

This HealthMap data set was last updated on 26th May 2012, and then manually checked for imported cases and cross-validated against dengue transmission extent based upon evidence consensus[2]. In total, the HealthMap data provided 1,622 new dengue occurrence data points in addition to those previously extracted as described in A2.2.

A.3.2: Assigning geo-positions to data from online informal data sources

Geo-positions for the HealthMap data were generated using a custom-built gazetteer, or geographic dictionary, of over 4,000 relevant phrases and place names and their corresponding geographic coordinates. The system uses a look-up tree algorithm that searches for matches between sequences of words in alert info text and sequences of words in the gazetteer. When a match is found, a set of rules are applied which attempt to determine the relevance of the place name to the outbreak that is being reported based on the position of the phrase in the report text. The gazetteer includes place names at a range of spatial resolutions (e.g. neighbourhoods, cities, provinces and countries) and uses certain phrases to trigger exclusion of a place name (e.g. Brazil nut). As with the formal occurrence records, all

informal occurrence records underwent temporal standardisation (A.5) and quality control (A.4).

A.4: Automatic validation and quality control

Geo-positioned data from both sources were entered into a bespoke PostgreSQL database that links disease data to spatial data in order to cross-check and validate data points. First, a raster distinguishing land from water was created at a 5km x 5km resolution and was used to ensure all disease occurrence points were positioned on a valid land pixel such that they could be used along with other covariate layers in our analysis (see Supplementary Information B). Any data that met the following criteria were excluded from the database:

1. Points found in countries or administrative divisions classified as unlikely to have a dengue occurrence based upon evidence consensus[2]. This classification was determined according to a qualitative evidence base that assessed consensus among a wide variety of evidence types on dengue presence or absence at a national and sometimes sub-national level[2]. This consensus ranged from complete agreement on absence (score of -100) to complete agreement on presence (100). We chose to exclude points in areas with scores of less than -25. This conservative criterion was intended to preserve points in areas of both proven dengue presence and uncertainty on dengue status.
2. Administrative division polygons having an area greater than 111km^2 (one decimal degree at the equator).

A.5 Temporal standardisation

The collected dengue occurrence data came in a variety of temporal forms. Some sources reported multiple cases in a single location throughout a year with no finer-scale temporal information. However, in other sources (particularly online sources), multiple cases in the same location throughout the year were presented as a new report each time subsequent transmission occurred. As a result we chose to define a single occurrence at a given unique location as one or more confirmed cases of dengue occurring within one calendar year (the finest temporal resolution available across all records). Our annual temporal standardization involved disaggregating occurrence points in the same location spanning multiple years into individual occurrences for each respective year, and aggregating occurrence points in the same location within the same year to form a single occurrence point attributed to that year. Occurrence points were considered overlapping if they lay on the same 5km x 5km pixel, or if they occupied the same lower administrative level unit for occurrence polygons. It should be noted that multiple different temporal definitions of an occurrence point were tested, such as no temporal standadisation or total temporal standardization where occurrence points mark where dengue has ever occurred. These variations were found to have a negligible impact on the resulting predictive maps and results.

A.6 Data summary

Once these procedures were complete, the final database used in subsequent modelling contained 8,309 occurrence observations (including 5,216 point locations and 3,093 small polygon centroids) covering a period from 1960 to 2012. Of these 8,309 occurrences, 7,050 were from the literature and case report database and 1,259 were from HealthMap. The number of occurrence points at each stage of quality control processing is summarised in figure SA1 below.

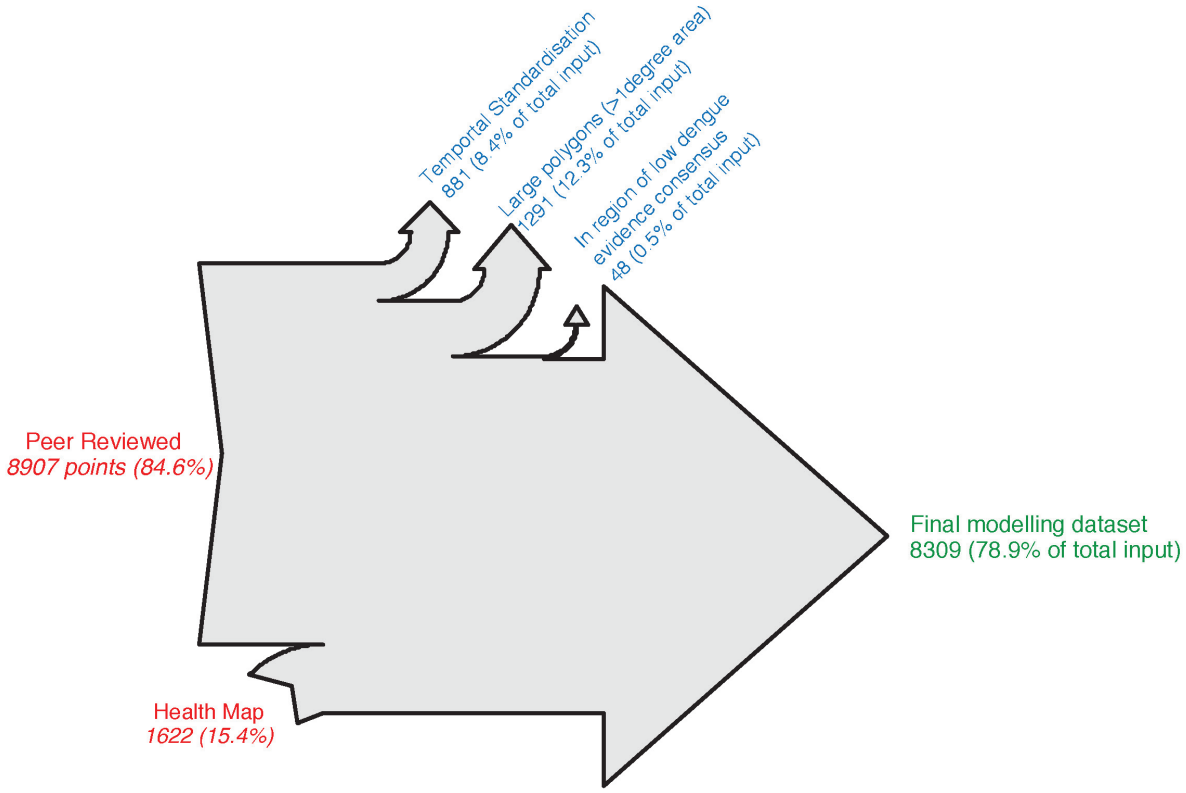

Figure SA1. Occurrence data processing summary. Text in red show the raw occurrence inputs, text in blue show the occurrence data lost through the stages of quality control and the text in green is the final dataset used in subsequent modelling.

Maps displaying the 8,309 locations are provided in Figures SA2-SA6 (polygon locations are represented by their centroids), the numbers of occurrence locations per year are shown in Figure SA7 and the temporal break-down by country and region are shown in Figures SA8-SA11.

The majority of occurrence records were sampled from the Americas (4215 - 50.7%) and Asia (3345 – 40.3%), with Africa (285 – 3.4%) and Oceania (464 – 5.6%) having fewer samples. The vast majority (86.5%) of the data is contemporary and sampled after 1990

(Figure SA7), but the temporal sampling density varies greatly by country (Figures SA8-SA11).

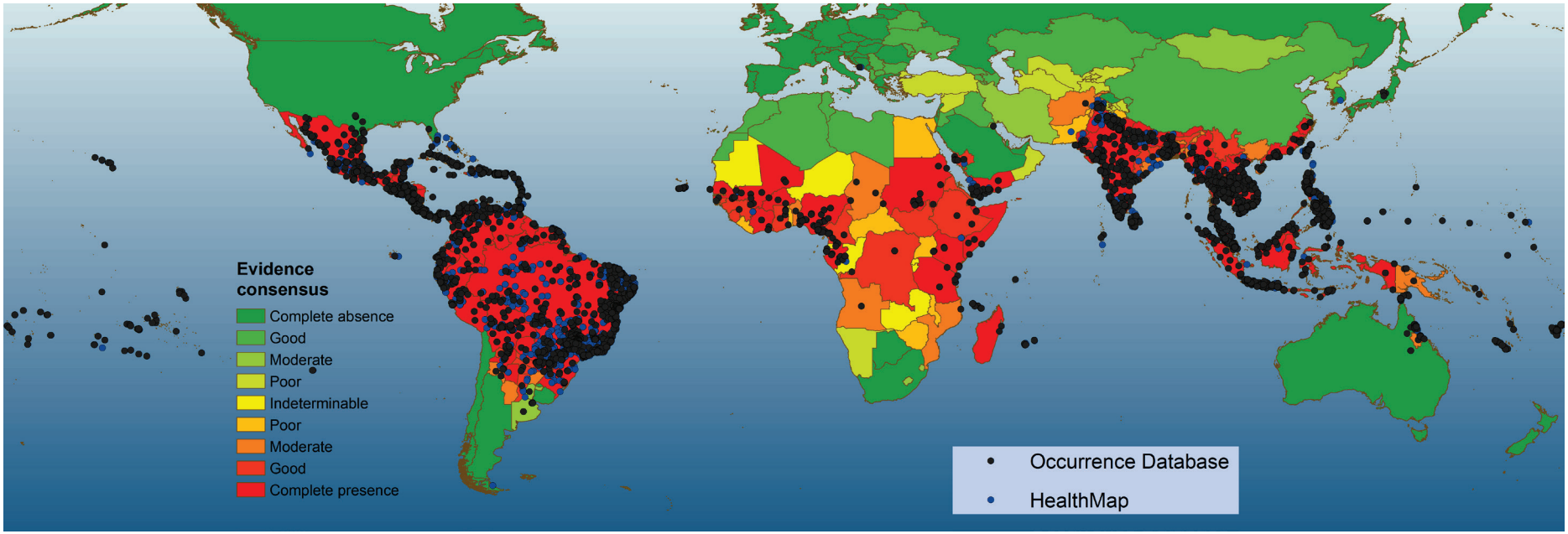

Figure SA2.Geographic locations of occurrence data globally. Country colouring is based on evidence-based consensus[2] with green representing a consensus on dengue absence and red a consensus on dengue presence.

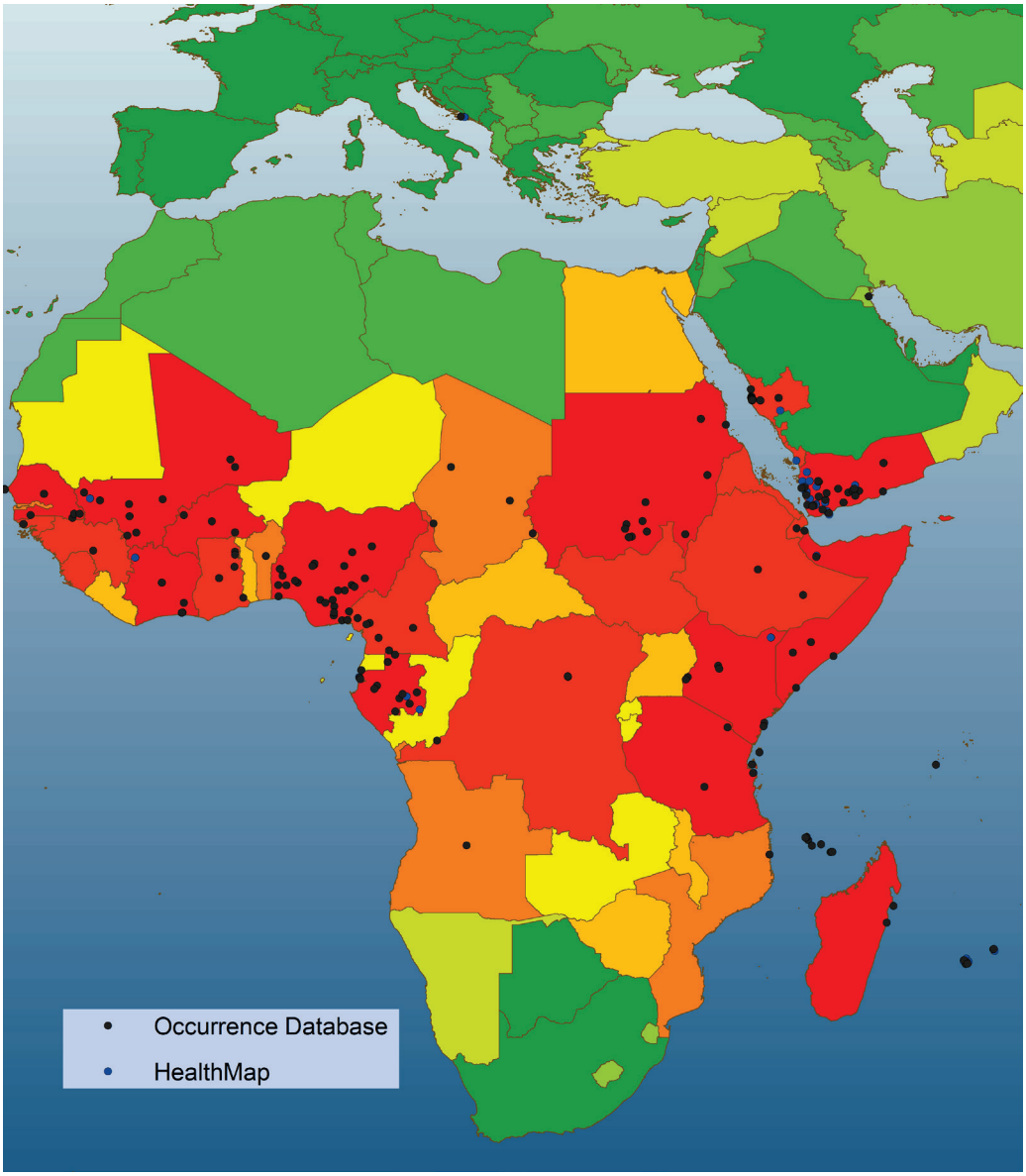

Figure SA3.Geographic locations of occurrence data in Africa and the Arabia Peninsula. Country colouring as per Figure SA2.

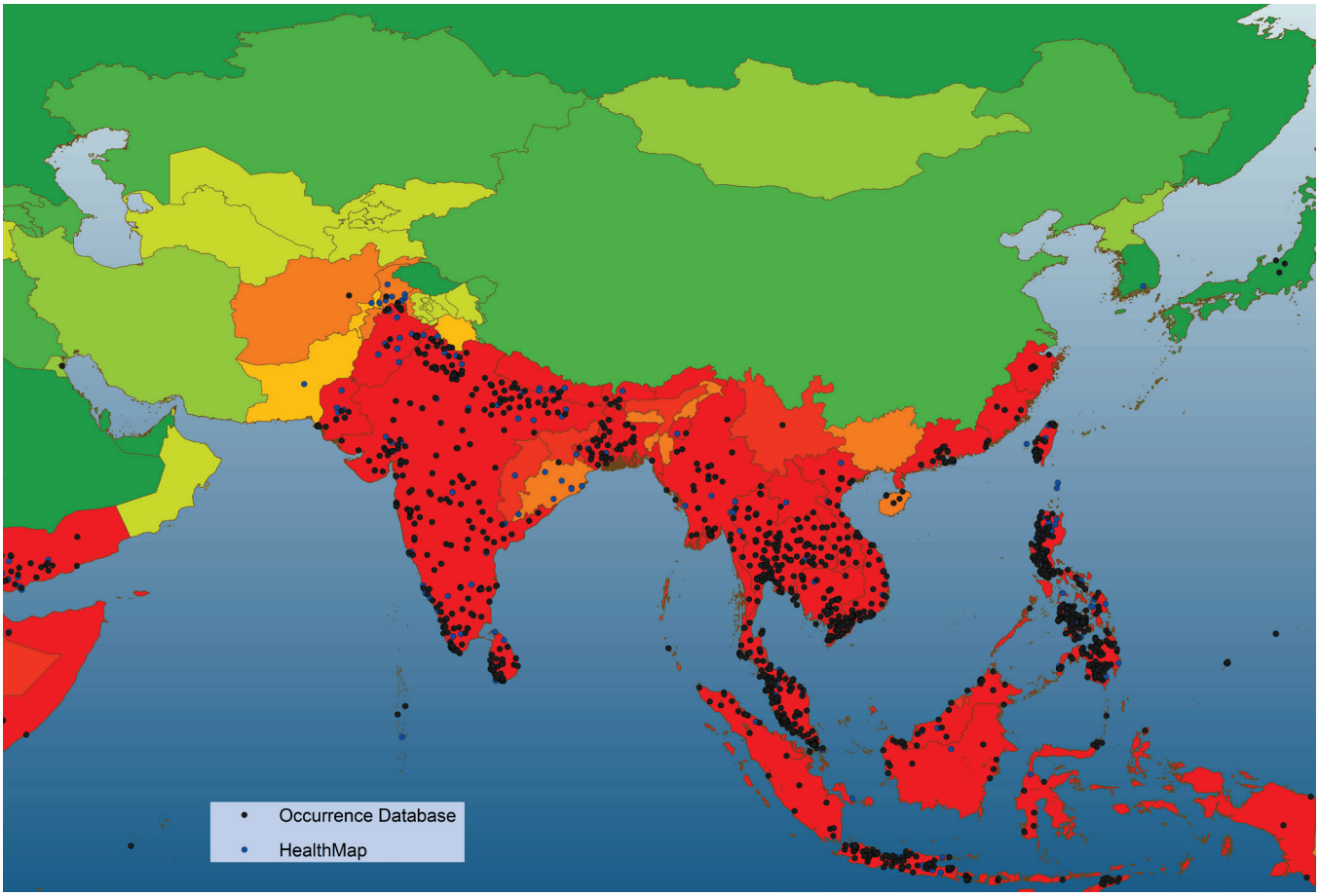

Figure SA4.Geographic locations of occurrence data in Asia. Country colouring as per Figure SA2.

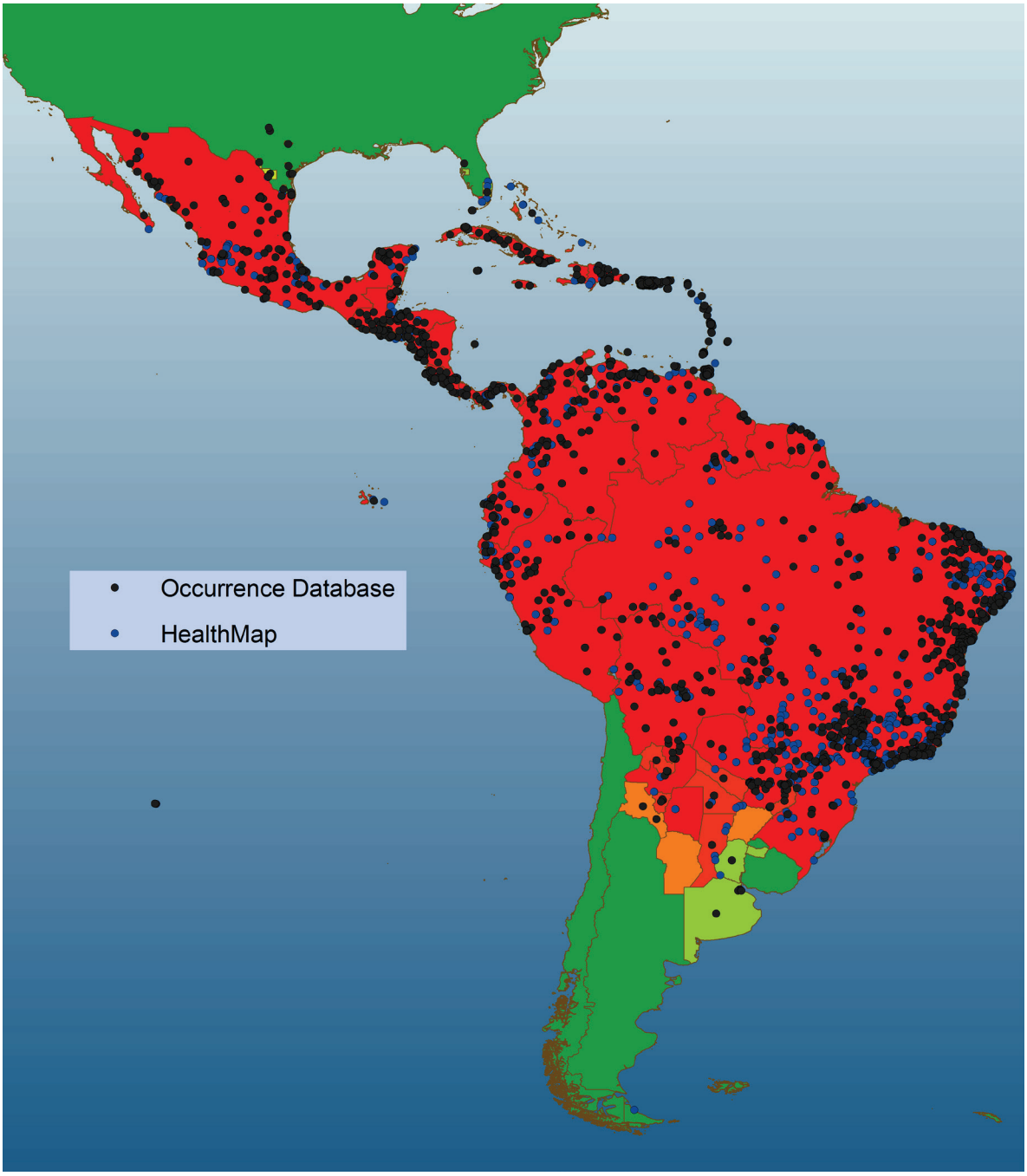

Figure SA5.Geographic locations of occurrence data in the Americas. Country colouring as per Figure SA2.

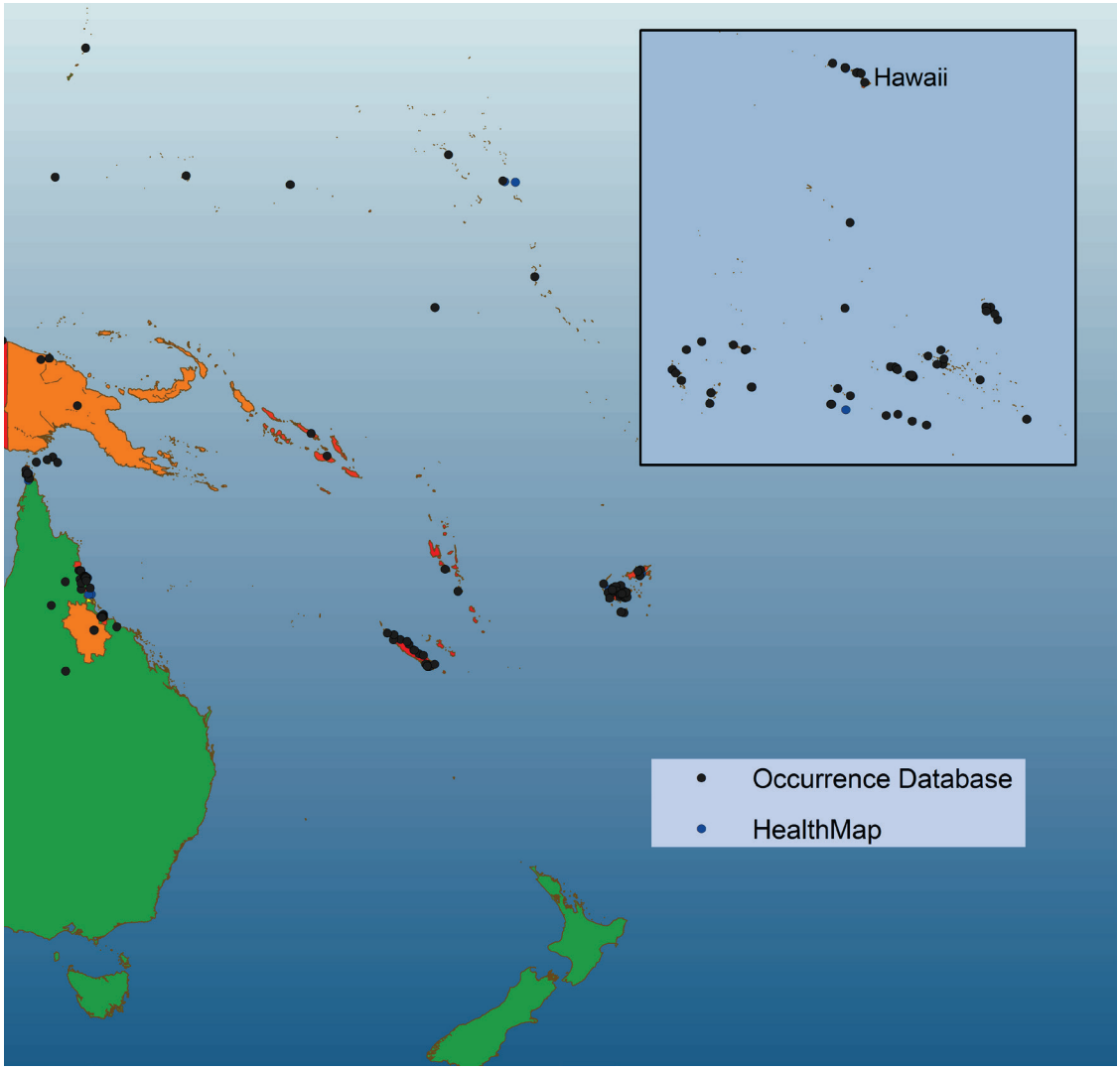

Figure SA6.Geographic locations of occurrence data in Australia and the Pacific.

Country colouring as per Figure SA2.

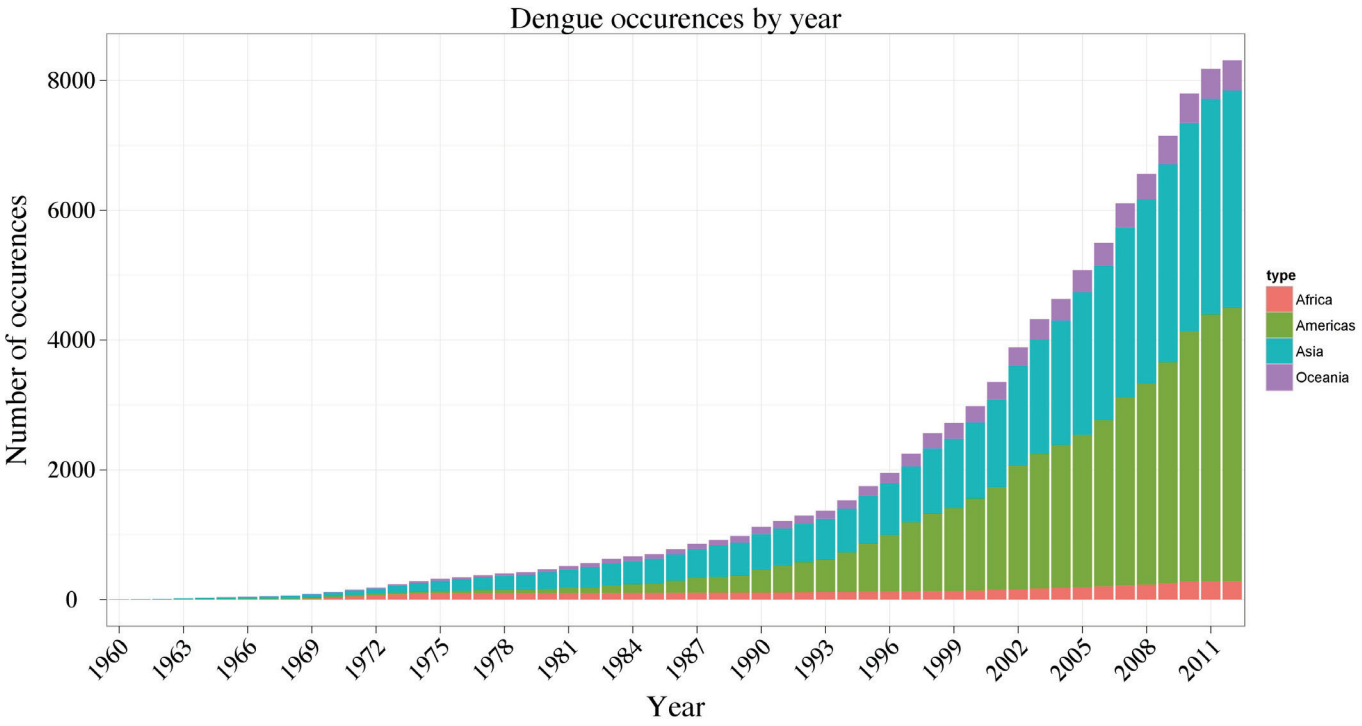

Figure SA7. Number of occurrence points per year globally. Bars are subdivided by continents Africa (red), Americas (green), Asia (blue) and Oceania (purple).

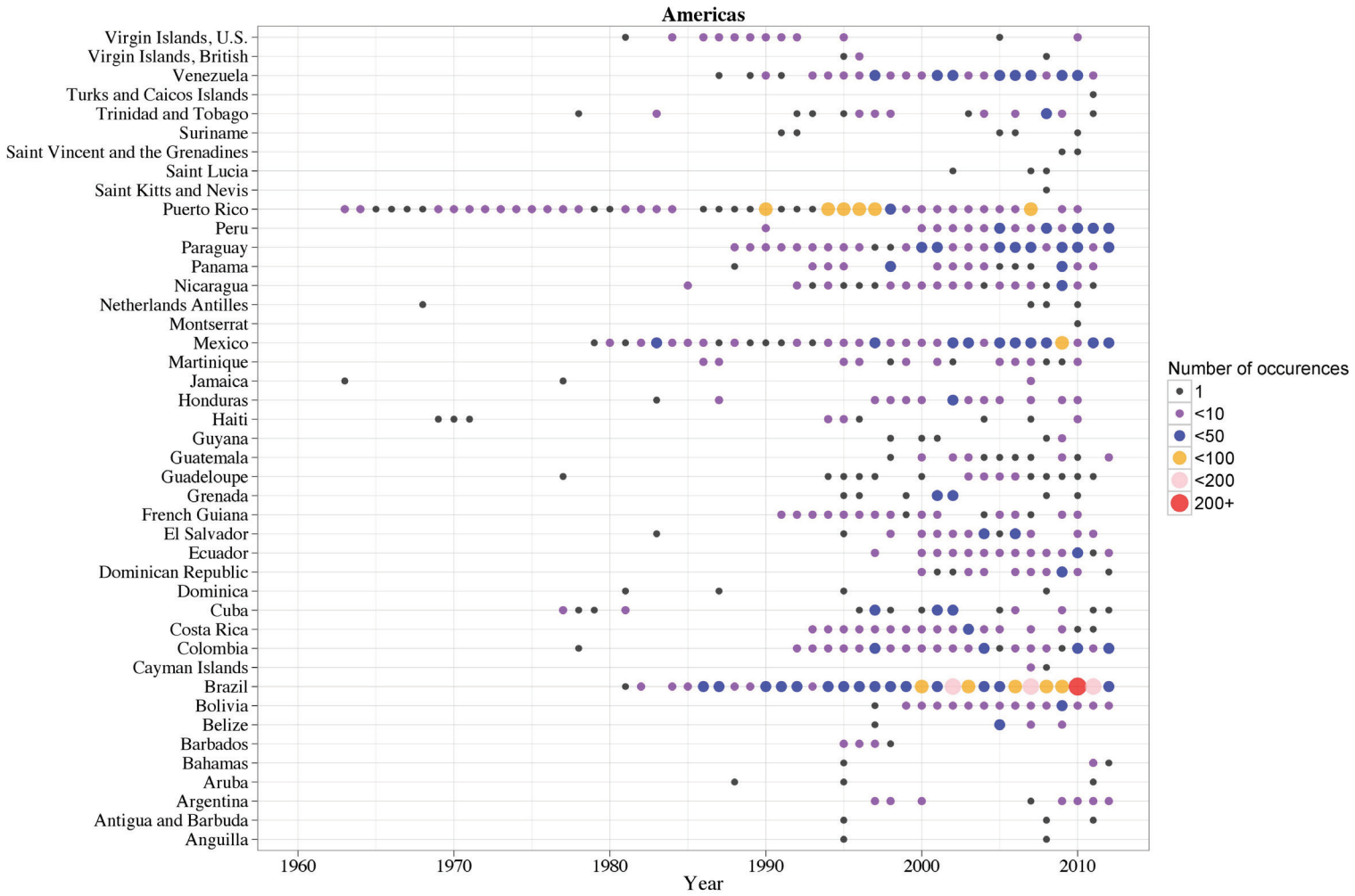

Figure SA8. Temporal breakdown of the number of occurrences per county in the Americas. Data point colour and size reflect the total number of occurrences at each time point.

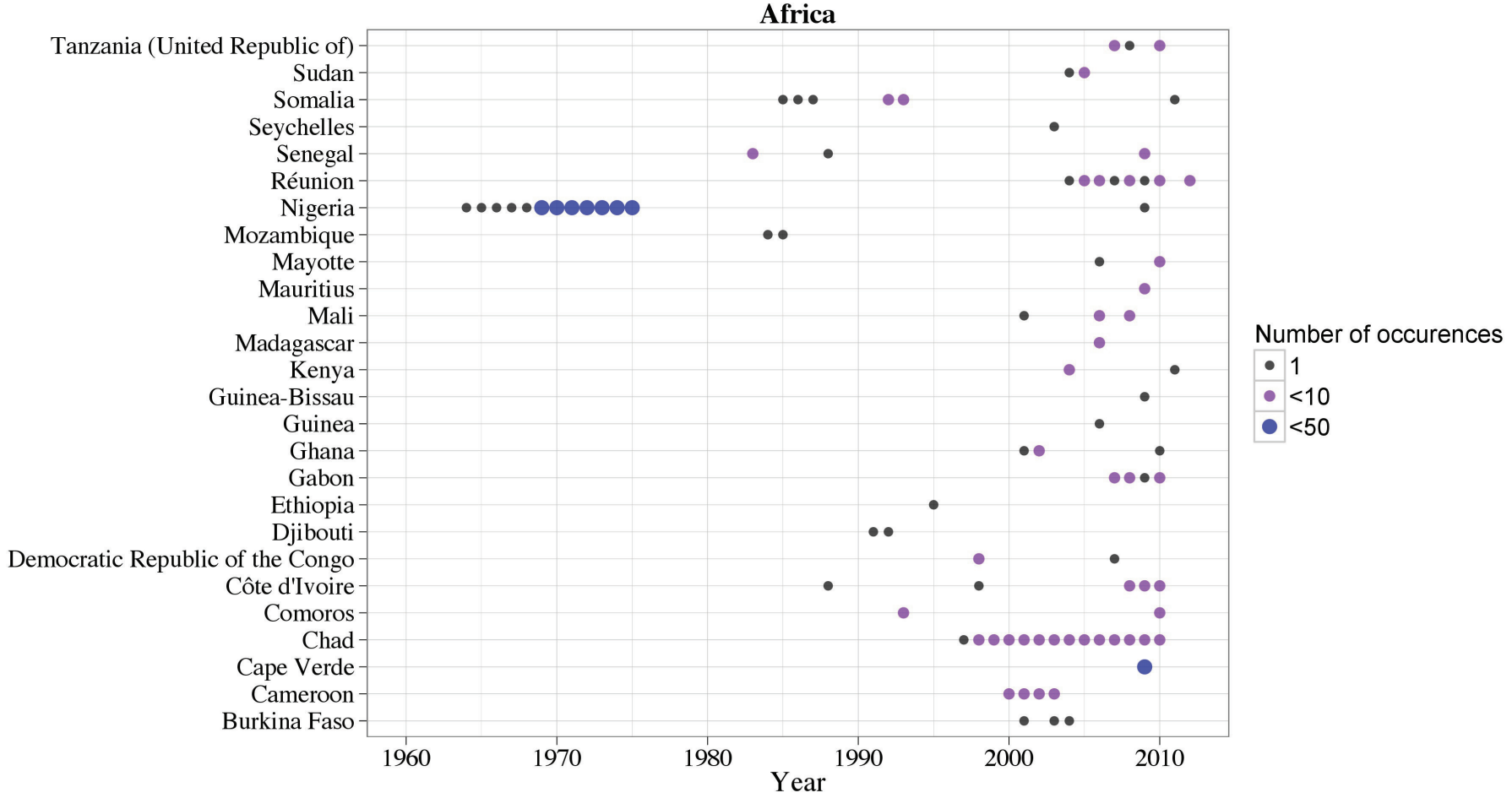

Figure SA9. Temporal breakdown of the number of occurrences per county in Africa. Data point colour and size reflect the total number of occurrences at each time point.

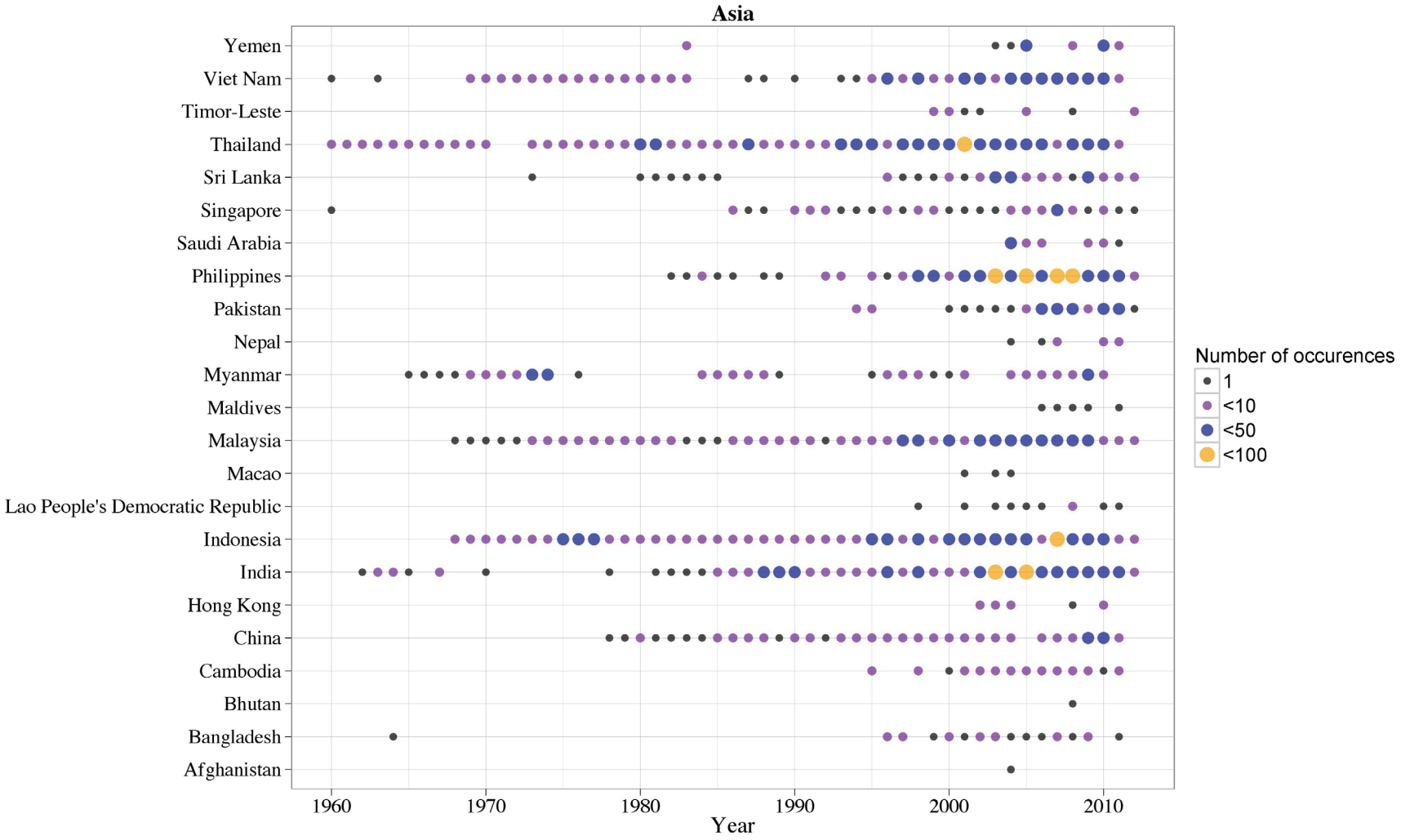

Figure SA10. Temporal breakdown of the number of occurrences per county in Asia. Data point colour and size reflect the total number of occurrences at each time point.

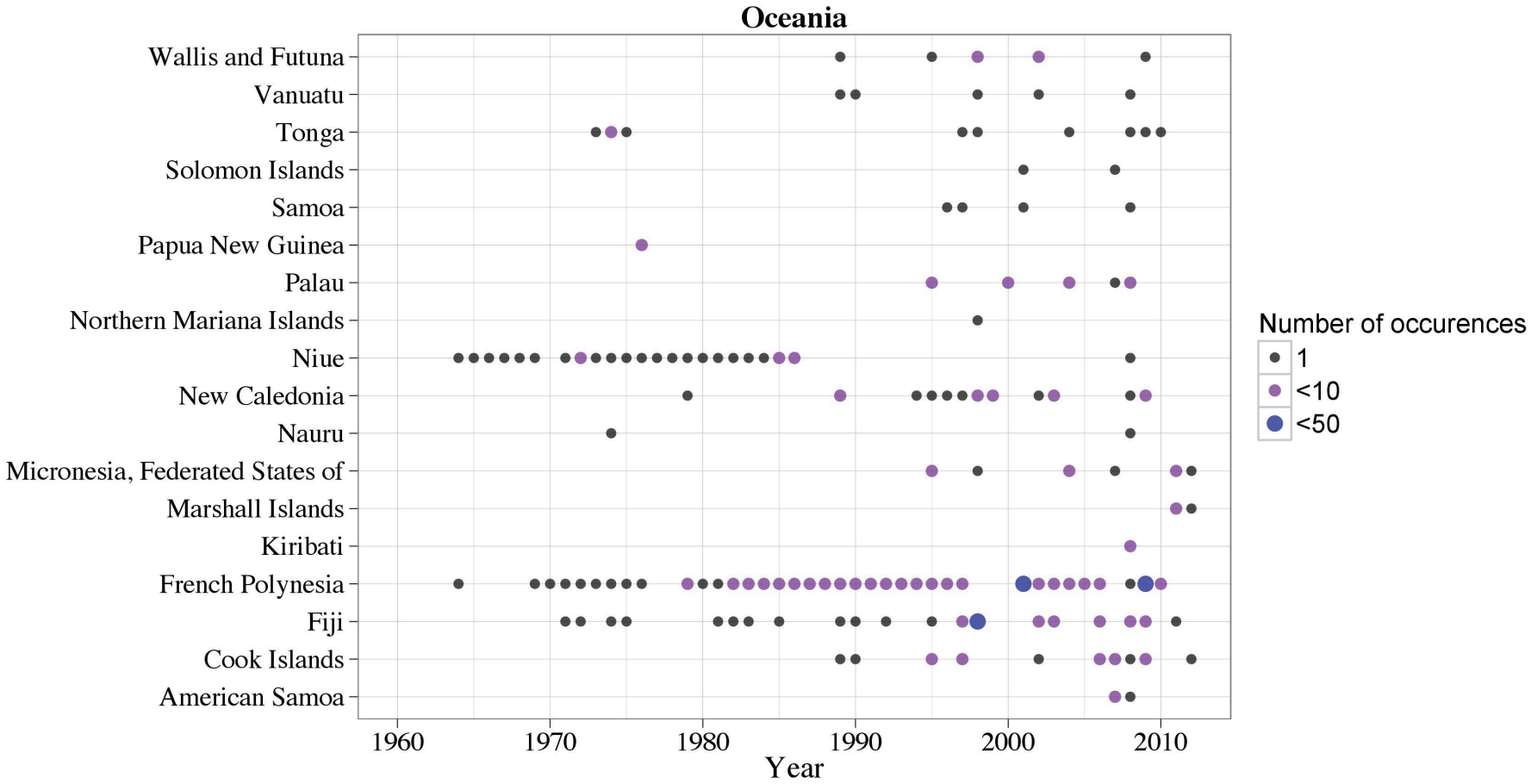

Figure SA11. Temporal breakdown of the number of occurrences per county in Oceania. Data point colour and size reflect the total number of occurrences at each time point.

B: Explanatory covariates

B.1: Overview

Dengue viruses are established in two habitats: the urban setting where humans and mosquitoes are the only known hosts, and forested areas where mosquito-borne viruses occur in nonhuman primates in a sylvatic cycle, with rare transmission from primates to humans[3-5]. Central to the global emergence of dengue virus has been the spread of its mosquito vectors. The primary vector of dengue virus is the highly domesticated, urban-adapted *Aedes aegypti*, found across tropical and subtropical latitudes[6]; however, other secondary vectors including *Aedes albopictus, Aedes polynesiensis,* and *Aedes scutellaris* can also transmit the virus. A complex interaction of factors influences the vector efficacy in virus transmission, with environmental factors such as precipitation, humidity, and temperature having been most-often incorporated into past efforts to model the distribution of dengue transmission[5-12]. However, multiple studies have emphasised the importance of socioeconomic factors in dengue transmission dynamics[10,13-15], such as the movement of mosquito vectors and viremic people[16], urban poverty and overcrowding, and poor public health infrastructure[3]. These factors have not yet been directly incorporated into global dengue distribution modelling until now.

For our model of the probability of occurrence of dengue virus, we used a suite of eight predictor variables. These covariates were chosen to reflect factors known or hypothesised to be ecologically relevant to dengue virus transmission dynamics, and for which it was feasible to collect data or derive proximate measures. The resulting set of covariates included (i) two precipitation variables interpolated from global meteorological stations, (ii) an index of temperature suitability for dengue transmission, (iii) a vegetation/moisture index, (iv) two

classes of urbanization, (v) an urban accessibility metric, and (vi) an indicator of relative poverty. All grids were standardised to ensure uniformity of land/water boundaries (as described in A1.5) and the same spatial resolution (5km x 5km). In this document, B.2 outlines our hypotheses underlying the choice of covariates, B.3 provides a detailed description of the sources for these covariates as well as how they were processed, B.4 describes the raster standardization process and B.5 and B.6 address statistical considerations.

B.2: Covariate selection

B.2.1: Climatic and environmental covariates

Precipitation:

Presence of static surface water in natural or man-made containers is a pre-requisite for *Aedes* oviposition and larval and pupal development. Despite *Aedes aegypti*'s principle larval habitats being man-made water storage containers[17], fine-scale temporal relationships between precipitation, vector abundance, and dengue incidence have been established in many locations[18-20]. These relationships are not universal, with dengue occurring in dry periods in some locations[21] and exhibiting varying patterns where two rainy seasons exist[22]. In general, however, there is evidence that areas with greater amounts of precipitation are associated with higher dengue infection risk[23,24].

Temperature suitability index:

As small-bodied ectotherms, *Aedes* mosquitoes' distribution, life cycle duration, survival, and behaviour are all dependent upon temperature[25,26]. Similarly, the extrinsic incubation period (EIP) of the dengue virus in the mosquito decreases at temperatures between 30 and 35^{o}C[27,28]. In combination, these relationships determine the occurrence of dengue in *Aedes* at temperatures above 18-20^{o}C[27,29]. A direct association has also been found between higher

temperatures and dengue incidence in humans[29-31]. Rather than simply including raw temperature values in our models, we incorporate these two principal temperature-dependent mechanisms (mosquito survival and virus incubation) in order to formulate a more biologically relevant covariate. An index of dengue-specific temperature suitability was thus created using a biological model, with temperature data as an input, to calculate the number of days per year that a given location on our global grid is suitable for dengue transmission[32-34]. This index is explained in greater detail in B.3.2.

Normalized difference vegetation index:

There is often a close association between local moisture supply, vegetation canopy development and abundance of breeding mosquitos [35], with previous studies highlighting the importance of moisture-related measures such as relative humidity to dengue occurrence[7]. Although resistant to desiccation, both *Aedes* eggs and adults require moisture to survive[36-40], with low dry season moisture levels substantially affecting *Aedes* mortality[40-42]. Vegetation canopy cover has previously been associated with higher *Aedes* larvae density[43-46] by reducing evaporation from containers, decreasing sub-canopy wind speed and protecting outdoor habitats from direct sunlight. To account for these factors, we used the normalized difference vegetation index (NDVI) as a potential indicator of the overall moisture availability and vegetation canopy cover at a given location.

B.2.2: Socio-economic covariates

Relative poverty indicator:

Several studies have linked poverty to dengue[47-50]. Typically, in both rural and urban settings, poorer areas are characterised by several factors that may favour higher dengue transmission. In many cases, relative poverty can be more indicative of economic disadvantage than

absolute poverty, as those living below a median or mean income threshold cannot derive the benefits of a sufficient material standard of living in relation to their circumstances. These standards comprise factors such as sustained vector control, access to principal health care services, manageable household sizes, basic sanitation, and reliable water supply. Lacking in any one of these standards may contribute to higher dengue transmission, and thus areas of greater relative poverty were hypothesised to exhibit a higher occurrence of dengue[51,52]. To account for relative poverty, we chose the finest geographic-scale data available for economic productivity and adjusted for purchasing power parity to reflect per-pixel relative poverty[53].

Urban accessibility:

At national and international spatial scales, individual human movements drive dengue virus introduction and reintroduction[54,55]. Indeed, the global spread of dengue virus in the past sixty years occurred through shipping routes and was characterised by periodic, large, spatial displacements[56]. Globalisation has further aided viral transmission by increasing the speed and frequency with which climatically suitable locations for dengue are connected [57,58]. Spread of dengue into new locations requires establishment of a competent local vector population, as the dispersal capabilities of individual mosquitoes are limited[59]. Conversely, movement of viremic humans occurs frequently, between a multitude of locations and at varying spatial scales. Therefore, human movement is the key facilitator of the spread of dengue virus at larger spatial scales[55], particularity in highly accessible, interconnected areas towards which people tend to gravitate. To simultaneously account for accessibility, patterns of human movement, and urban gravitation, we use the time required to travel from a given geographic location to a large city (minimum population 50,000) via land or water-based transportation networks[60,61].

Urbanisation:

While dengue transmission has been documented in both rural and urban settings [62], urban environments are characterised by many factors that are favourable for dengue transmission. These typically include population growth, a high abundance of vector breeding sites resulting from poor hygiene, inadequate housing quality, and minimal environmental management practices. Consequently, a high proportion of people in urban environments are brought into contact with the *Aedes* vector, resulting in a disproportionate degree of new and sustained dengue transmission compared to rural localities[63,64]. Peri-urban environments also constitute a large proportion of the area where dengue is found in tropical and sub-tropical regions. Unplanned settlement, overcrowding, and routine household water storage behaviour in these settings all combine to produce higher likelihood of vector abundance and viral transmission[65-67]. We created a categorical variable to differentiate between urban, peri-urban, or rural areas by supplementing the 2010 Global Rural Urban Mapping Project (GRUMP) urban and rural categories with land-cover classes to further distinguish peri-urban extents[68].

B.3: Covariate sources

B.3.1: WorldClim database - precipitation

The WorldClim database (www.wordclim.org) consists of a freely available set of global climate data at a 1km × 1km spatial resolution which was compiled using weather data collected from world-wide weather stations[69]. The data spans the period 1950-2000 and describes monthly averages of precipitation during this period. From these data, interpolated global climate surfaces were produced using ANUSPLIN-SPLINA software[70]. The result is a composite data set encompassing multiple time intervals, from which we extracted information about seasonal and inter-annual variation in precipitation patterns for each

gridded cell of our interpolated surfaces with temporal Fourier analysis (TFA) to generate the minimum and maximum monthly precipitation averages for the entire time series[71,72].

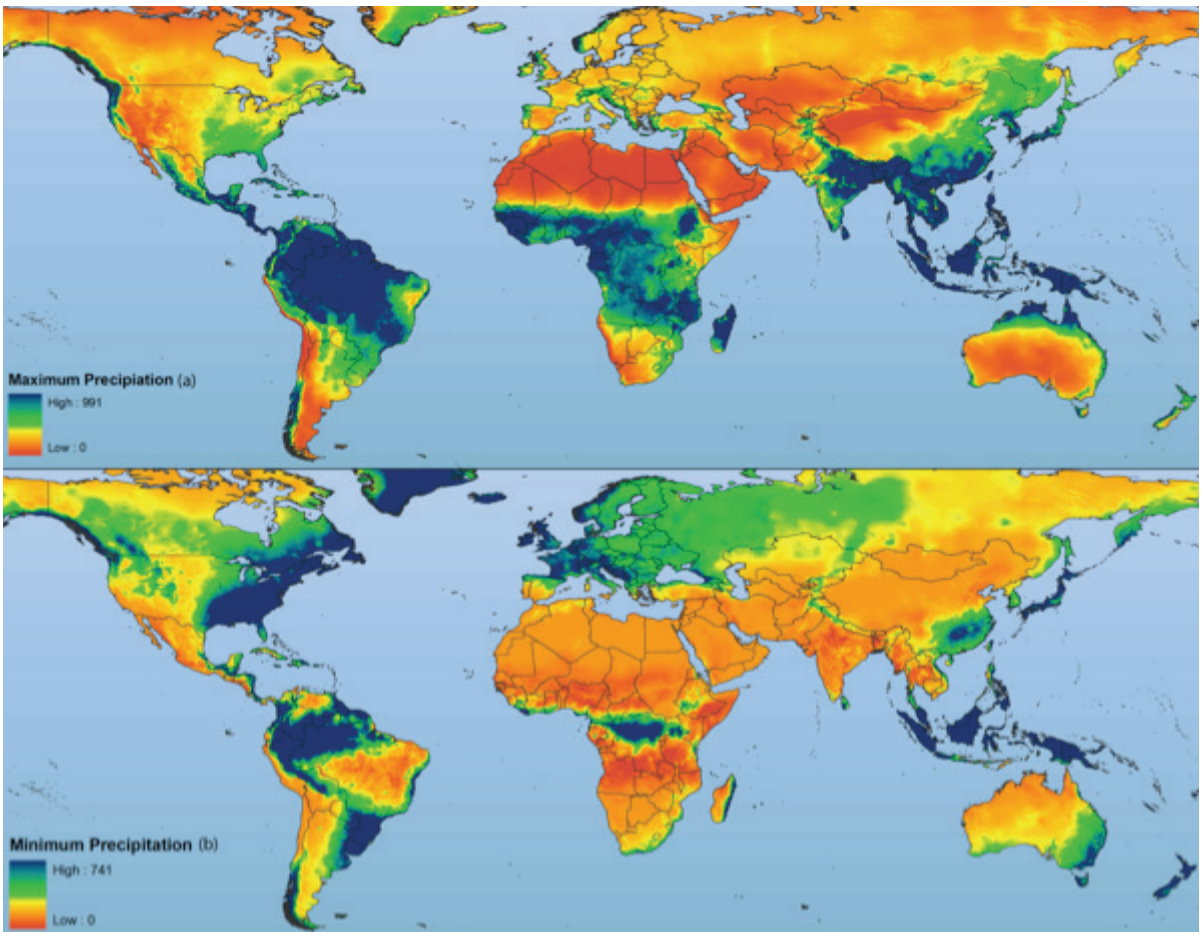

Figure SB1. WorldClim TFA maximum (a) and minimum (b) precipitation averages (mm).

B.3.2: Temperature Suitability

We developed a biological model for suitability of dengue virus transmission across intra-annual temperature cycles. This measure of days of suitable transmission was calculated for every 1km x 1 km pixel globally in an approach similar to that devised for applications in malaria research[34]. This dynamic model incorporates the effects of continuously changing temperature regimes on vector and virus survival. Temperature suitability, then, is an index Z which is proportional to vectorial capacity V, or the daily rate at which future infectious bites

will arise from one infective case. *V* depends on two principal temperature-dependent mechanisms of the dengue transmission cycle: (i) the life span of the *Aedes* vector and (ii) the duration of the extrinsic incubation period (EIP) of the dengue virus[32,33]. The quantitative relationships entered into our models were defined following the findings of Focks *et al*. (1993), with a constraint placed on EIP such that it may never exceed the maximum vector lifespan. Using WorldClim temperature data for 1950-2000 which was TFA-processed in the same manner as the precipitation data (S2.3.1), daily temperature estimates and a sinusoidal diurnal cycle for all 365 calendar days were interpolated from the synoptic monthly values (minimum, maximum, mean) for each pixel using a cubic spline as in Gething *et al*. 2011[73,74]. We then computed *Z* for all days within every pixel; all values of *Z* greater than 0 indicated suitability for dengue transmission. The total number of days *Z* was greater than 0 was summed for every pixel throughout one year (ranging from 0 to 365). This was then rescaled from 0 to 1 to create the final temperature suitability index included in our dengue occurrence distribution models. Pixels where $Z=0$ were considered permanently unable to support dengue transmission and were consequently designated as having a zero probability of occurrence in our final risk maps.

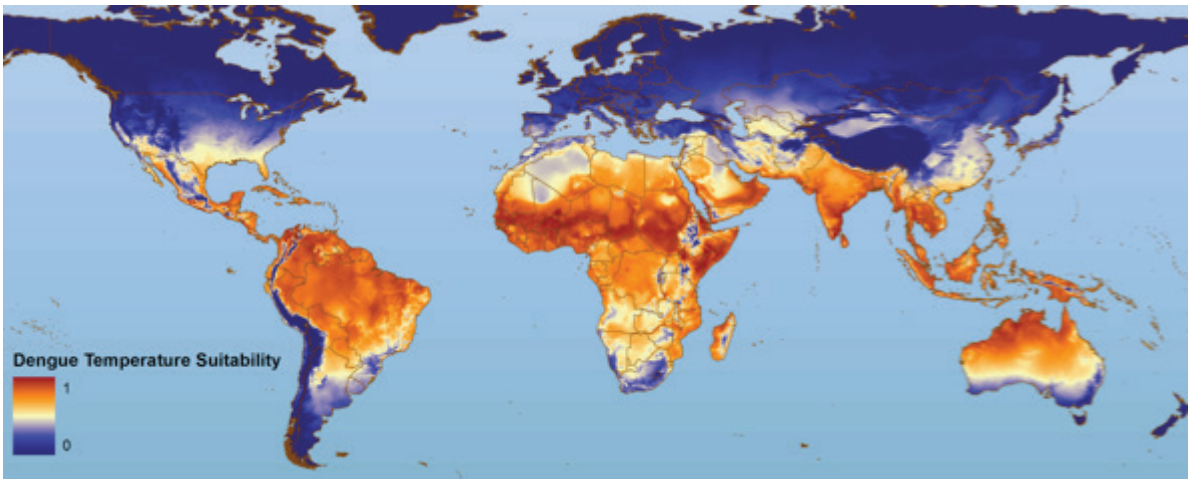

Figure SB2. Temperature suitability index. Created from a biological model for the suitability (zero to one with one as most suitable) of annual temperature patterns for dengue transmission.

B.3.3: Advanced Very High Resolution Radiometer - NDVI

The Advanced Very High Resolution Radiometer (AVHRR) 8km × 8km products are available over a 20-year time series, and a limited series of 1km × 1km resolution data are available for April to December 1992; January to September 1993; February to December 1995 and January to April 1996. We used the AVHRR NDVI product which numerically specified the level of green, photosynthesizing, and therefore active, vegetation derived from the spectral reflectance of AVHRR channels 1 and 2 (visible red and near infrared wavelength, respectively)[75,76]. From a composite data set encompassing multiple time intervals, we extracted information about average temporal variation in NDVI patterns for each gridded cell of our interpolated surfaces by TFA.

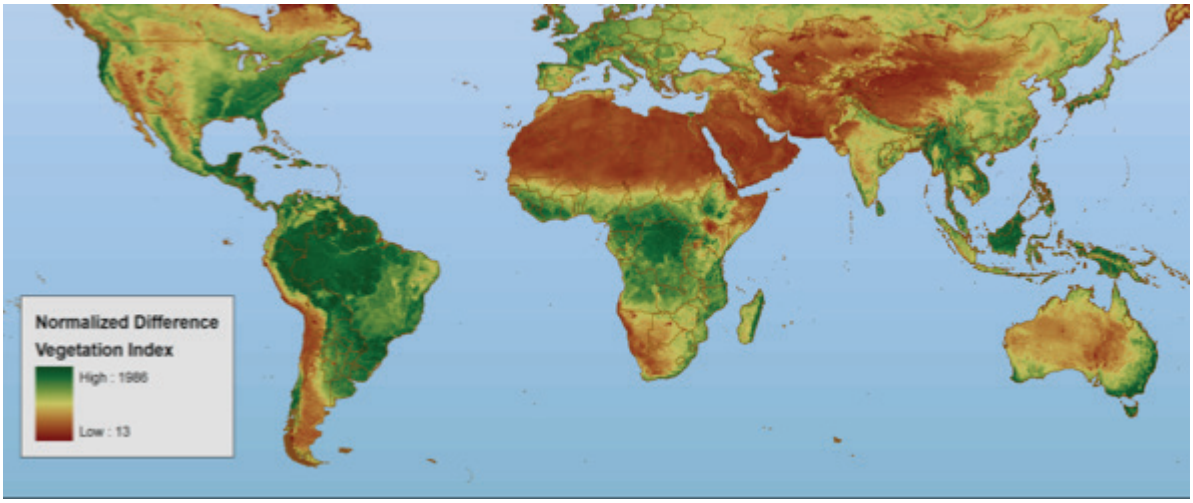

Figure SB3. AVHRR TFA mean normalised difference vegetation index.

B.3.4: Global Rural Urban Mapping Project

The Global Rural Urban Mapping Project Urban Extents (GRUMP-UE) surface is created principally using night-time lights satellite imagery, supplemented with data derived from tactical pilotage charts and known settlement points[68,77,78]. Through using satellite night-time lights as the basis for mapping urban areas, GRUMP-UE has been shown to overestimate urban extents due to the "overglow" effects seen in such imagery[79], resulting in the inclusion of less intensely urban, or "peri-urban," areas. Previous work has shown that areas where population density is greater than or equal to 1000 people per km^2 are indicative of intense urbanization, thus providing a suitable threshold for distinguishing urban from peri-urban areas[80-83]. This was implemented using the Gridded Population of the World version 3 (GPW3)[68] population density database projected for 2010[80]. This database is derived from the most recently available national censuses and other demographic data, resolved at the highest possible administrative boundary level, and area-weighted[83] to a 5km × 5km spatial resolution grid. The final result was two variables for our distribution modelling, the first which identified a pixel as urban or otherwise, and the second which identified a pixel as peri-urban or otherwise.

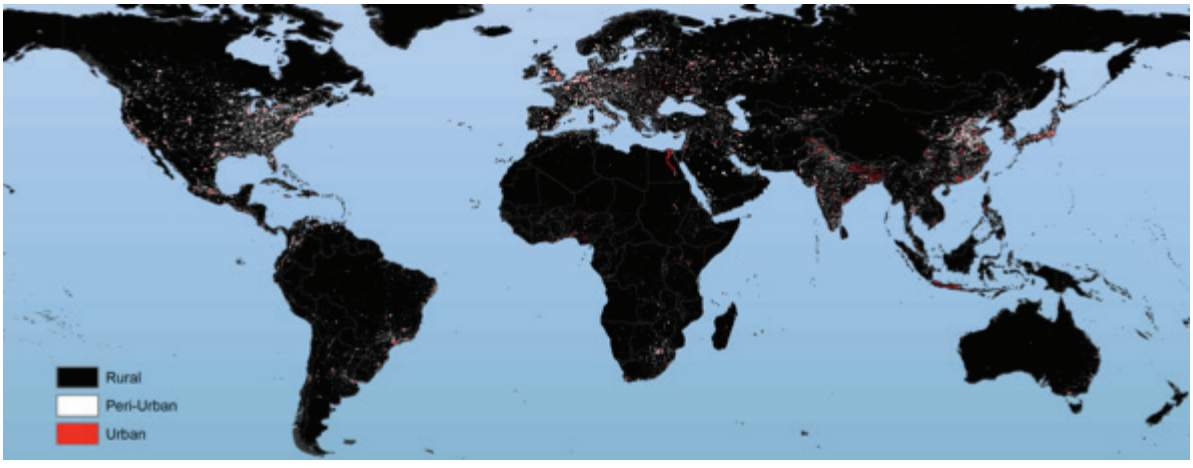

Figure SB4. GRUMP urban and peri-urban categorical classification.

B.3.5: Urban Accessibility

The urban accessibility data set was obtained from the European Commission Joint Research Centre Global Environment Monitoring Unit (JRC)[61].This 1km × 1km-resolution database defines the travel time to a city of 50,000 people or more in the year 2000 using land- or navigable water- based transportation methods. This is computed using a friction-of-distance algorithm which computes the 'cost' in time of travelling between two locations on a regular raster grid[60,61]. It is derived from several spatial data sets representing roads, terrain, shipping lanes, land cover, political boundaries, and any other geographic features that should be considered when estimating the travel time to target locations. Consequently, the target locations were cities with a population of 50,000 people or more in the year 2000 based upon the Global Rural Urban Mapping Project (GRUMPv1) human settlements database[84].

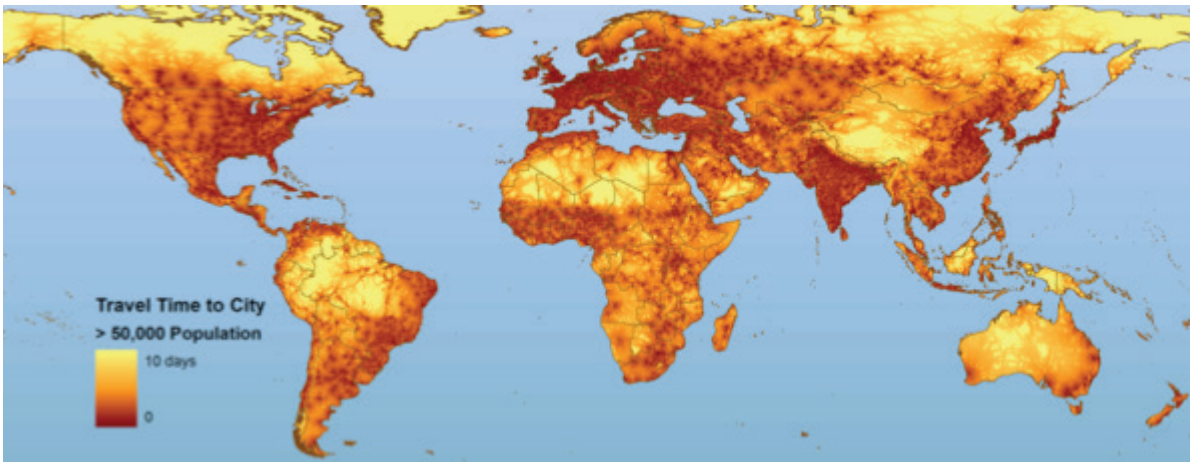

Figure SB5. JRC urban accessibility. Defines the travel time to nearest city with a population of 50,000 or more by land or water based transportation.

B.3.6: Relative Poverty

Most global measures of economic activity are time series measured at the national level, providing a very limited number of observations at enormously different geographic scales. The G-Econ database (http://www. gecon.yale.edu), however, takes economic estimates

from the lowest possible administrative subdivision for which data are available and then spatially rescales[85] these data to provide a global grid of economic activity at a 111km x 111km (1 degree at the equator) resolution. A detailed explanation of the spatial rescaling methodology can be found in Nordhaus *et al.* (2006; 2008). One or more of four major sources of economic data for each administrative unit are utilised in order to create the database: (i) gross regional product (ii) regional income by industry (iii) regional employment by industry and (iv) regional urban and rural population or employment along with aggregated sector data on agricultural and non-agricultural incomes[85].

The result is a measure of gross cell product (GCP) for each 1 degree x 1 degree cell globally, with the same conceptual basis as gross domestic product (GDP), referring to the total market value of all final goods and services produced within one year, and is generally thought of as an indicator of the overall standard of living in a given area. In some cases, the original G-Econ database contained multiple entries for a single cell. When this was the case, we chose the value derived from the best-quality information (indicated by a "quality" field) and/or that which was most recently entered into the database. We then adjusted the GCP measures for purchasing power parity (PPP) in U.S. dollars for the years 1990,1995,2000, and 2005, using national aggregates estimated by the World Bank[86] and computed the mean across all years for each gridded cell globally. This PPP-adjusted measure of GCP served as the indicator of relative poverty used as a covariate in our model.

Figure SB6. G-Econ gross cell product in U.S. dollars. Values averaged over years 1990, 1995, 2000 and 2005 and adjusted for purchasing power parity.

B.4: Raster Standardisation

As detailed in this document, the original data sources for our covariates came in a variety of formats, with varying spatial resolutions. Additionally, the land-water boundaries inevitably differed slightly between data sets obtained from different sources, such that the precise definition of coastlines and the inclusion or exclusion of small islands and peninsulas was not consistent. These factors precluded immediate use of these data in a single spatial model. To overcome these incompatibilities and generate a fully standardised suite of input grids, we derived a standard geographic template around which all grids were processed. This template was implemented as follows: (i) input data sources were re-projected, where necessary, using a standardised equirectangular Plate Carrée projection under the World Geodetic System 1984 coordinate system; (ii) where input grids were defined at spatial resolutions other than 5km x 5 km, they were aggregated or disaggregated to this resolution using bilinear interpolation; (iii) grids were either extended or clipped to match a standardised extent spanning -180° east to 180° west, and from 85° north to 60° south; (iv) alignment to a standardised land-water boundary raster mask (see A.5) was performed using nearest

neighbor interpolation, ensuring a consistent coastline definition.

B.5 Covariate Extraction

Extraction of covariate values from occurrence degree coordinates was performed differently for occurrence points and polygons. For a given coordinate, points were assigned the covariate value of the pixel at that location. For polygon occurrences, covariate values were averaged across the all pixels contained in the polygon.

B.6: Multicollinearity

Multicollinearity occurs when two or more predictor variables in a statistical model are linearly related. Multicollinearity can lead to unstable parameter estimates and inflated standard errors on estimates[87]. As a rule of thumb, multicollinearity results in variance inflation when covariate variables have correlation coefficients[87] of $|r| > 0.7$. Pairwise correlation coefficients for all our covariate variables are well within this commonly used threshold and therefore multicollinearity is unlikely to affect our analysis (Table T1).

Table T1. Pairwise correlation matrix between covariate variables. Symmetric matrix with diagonal elements having a correlation of 1.TS (temperature suitability), RP (relative poverty), UA (urban accessibility), U (urban), PU (peri-urban), Pmin (precipitation minimum), Pmax (precipitation max), NDVI (normalized difference vegetation index).

	TS	RP	UA	U	PU	Pmin	Pmax	NDVI
TS		-0.120	-0.017	0.141	-0.076	0.002	0.268	-0.226
RP			-0.124	0.213	0.048	0.098	-0.056	-0.035
UA				-0.161	-0.018	0.143	0.026	-0.031
U					-0.463	-0.031	0.063	-0.500
PU						0.044	-0.115	0.186
Pmin							0.194	0.201
Pmax								0.069
NDVI								

C: Predicting probability of dengue transmission using Boosted Regression Trees

C.1: Overview

Knowledge about the geographical distribution of diseases is central to the planning, implementation, and monitoring of control programmes, and also underpins approaches to future risk prediction and mitC.3igation. However, in most cases, detailed data on disease distributions are not available and collecting such data is costly and labour-intensive[88]. Consequently, there has been an increased interest in developing predictive modelling approaches to estimate disease distribution patterns when dealing with incomplete data[89]. Such approaches vary according to the nature of available data (i.e., whether it relates to disease prevalence, incidence, or occurrence) and the locational specificity of the available data (i.e., whether precise point locations or administrative regions). Particularly extensive population-representative surveillance data on prevalence or incidence are rare for infectious diseases, especially those which are more neglected from a public health perspective. More commonly, the only data available for mapping these diseases are observations of their occurrence in different locations, without corresponding information about where they are known to be absent or less prevalent. Generating disease maps from occurrence point data is thus similar to estimating species distributions, which characterise habitats suitable for a given species (niche modelling)[90,91] based on geo-referenced collection locations. In the context of disease mapping, the aim is to determine habitat suitability for the persistence of a given disease agent and its transmission vectors at sufficient levels to result in human cases. This suitability may be determined based upon the climatic, ecological, and socioeconomic characteristics of those locations where the disease has been reported.

Nearly all species distribution models (SDMs) comprise a decision boundary in a multidimensional space in which each dimension represents a different environmental variable (for example average temperature, rainfall patterns, degree of urbanization, etc.). The modelling objective is to characterise the subset of this environmental space where the species or disease occurs, which then allows the suitability of all other locations across a map to be assessed. This results in estimates of the probability of the species or disease occurring in these other locations. Whilst a diverse array of models has been developed, they generally differ only in the way this multidimensional characterisation or 'response' is achieved. Two broad classes of SDMs can be distinguished: profile and classification models.

Profile approaches require only presence data to determine habitat suitability. The BIOCLIM[92] model is a popular example that captures the environmental niche, or 'profile,' of a species by creating a rectangular envelope in the multidimensional space, defining the limits of the species' spatial range based upon the most extreme (minimum and maximum) values of each environmental variable in the locations where it has been observed. Whilst the reliance only on presence data only is advantageous, these envelopes are simplistic in that they cannot differentiate between varying densities of occurrence records within the defined limits and accordingly do not allow for a probabilistic representation of the predicted species or disease distribution.

Classification techniques are derived from statistical and machine learning algorithms and have been shown to have a greater predictive capacity than the simpler profile methods[93-97]. One well-known example of a classification approach is the generalised linear model (GLM)[98], which represents multivariate space using linear parametric terms, such as a

combination of quadratic or cubic equations. More elaborate regression techniques have been developed to overcome the limited flexibility of parametric forms and allow the modelling of complex ecological response shapes[99]. These include fitting non-linear functions either additively (e.g. generalised additive models, GAMs[100]) or piecewise (e.g. multiple additive regression splines, MARS[101]), or by recursively partitioning the environmental space into a large number of subsets within which separate regression models are fitted and then recombined to give a complex final response (regression trees[102]). Maxent[103] is a popular machine learning approach that estimates disease distributions by finding the distribution of maximum entropy[104,105]: the simplest possible distribution that is consistent with the mean and variance of the observed distribution.

The flexibility of a model to fit complex environmental responses must be weighed against the danger of overfitting, where the model is tuned to noise present in the data as well as the underlying signal, rendering the model less accurate in prediction. An approach that has drawn significant research attention in recent years is boosted regression trees (BRT)[100,106-108], which combine the complex fitting capability of a regression tree with boosting, a variance reduction technique that consists of iterative improvements to the model obtained by importance sampling. Boosting allows fine tuning of the overall model fit whilst reducing the variance of predictions[107]. By including a cross-validation procedure at each iterative step (whereby model performance is evaluated against a randomly held-out subset), BRTs are also adept at avoiding over-fitting[106,108].

A comprehensive comparison of 16 modelling methods found that machine learning methods tend to out-perform the more traditional regression approaches with regards to prediction performance; among these Maxent and BRTs were the best[94]. Whilst broadly comparable,

BRTs were marginally superior at capturing complex responses and have achieved greater specificities in predicting areas of specified absence[109]. We therefore elected to adopt a BRT approach in this study.

Like all classification methods, BRTs require input data for both presence and absence of the species or disease in question[94,110,111]. This requirement arises from the conceptual objective of modelling the environmental characteristics of areas associated with observed presences relative to the entire available environment. Information on the available environment is provided by a sample of background points (or pseudo-data) from the study region. Background points do not define the distribution of disease absence; rather, they provide a sample set of conditions in places where the disease has not yet been observed. Consequently it is critical that generation of the background pseudo-data is informed by a good understanding of the factors shaping the geographic distribution of the presence data[95]. Consideration must, therefore, be taken when selecting the amount, location and geographical configuration of pseudo-data.

In the following sections we provide both a conceptual and a technical description of our BRT model structure and details of its implementation. We then explain our protocol for sampling data from a contrast class and describe our ensemble analysis aimed at providing robust final output predictions irrespective of the different modelling decisions.

C.2: Boosted Regression Trees

C.2.1: Regression trees and boosting: a conceptual description

BRTs combine regression or decision trees with "boosting". The regression tree component builds a set of decision rules on the predictor covariates. These rules are constructed by

recursively partitioning the data into successively smaller groups using binary splits. Splits for all of the predictors are repeatedly applied to their own output until the best split is chosen[108,112]. For regression trees, the best split is that which maximises the homogeneity of the two resulting groups with respect to the response variable[113]. The output is a decision tree with the branches determined by the splitting rules and a series of terminal nodes ("leaves") that contain the mean response. To reduce variance we used a boosting meta-algorithm. In the context of regression trees, boosting is a form of functional gradient descent[113], which seeks to minimise a loss function (in our case the residual deviance) by adding, at each step, a new tree that best reduces, or steps down, the gradient of the loss function. Therefore, in the combined BRT procedure, a regression tree is first fitted to minimise loss. Then boosting is performed in a forward stagewise manner to further minimise residual variation in the response. The final model is a linear combination of many trees that can be thought of as an additive model in which each term is a tree[113]. Forward stagewise fitting also makes it easy to use cross-validation to optimise the number of trees and prevent over-fitting. Formal mathematical descriptions of classification and regression trees can be found here[102], gradient boosting can be found here[107] and boosted regression trees can be found here[100]

C.2.2: BRT parameter selection

The BRT approach requires the following parameters to be determined or specified: (i) the loss function; (ii) the number of tree/iterations in the stagewise additive model (*m)*; (iii) the interaction depth *K*; (iv) the learning rate ν and (v) the stochastic subsampling proportion π. For the loss function (i) we chose a binomial loss function:

$$\ell(y, f(x)) = \log\left(1 + \exp(-2yf(x))\right) \quad (1)$$

This function was chosen not only for applicability to binary data but also for robustness[106,114]. For parameters 2-5, we follow Elith et al. (2008)[106] in setting the interaction depth *K* equal to 4, the stochastic subsampling proportion π equal to 0.75, and the learning rate *v* equal to 1% (a slow rate chosen for optimal performance). Determining the optimal number of trees is important, as the model can continue to add trees until over-fitting occurs and predictive performance is reduced. The optimal number of trees was found with 10 fold cross-validation using the methods of Elith et al.(2008)[106].

C.2.3: Summarising the BRT model

BRTs produce an ensemble of thousands of regression trees. To visualise these ensembles we constructed partial dependence plots and estimated the relative importance of covariates as follows[107].

Partial dependence plots:

The partial dependence functions[107] can be used to visualise dependencies between the response and the covariates. When plotted, the partial dependence function shows the marginal effect of each covariate on the response after averaging the effects of all other covariates. For the BRT, this integral can be approximated using the weighted tree traversal method[107] which uses the ensemble of regression trees and calculates the proportion of data that fall in the different terminal nodes for each covariate.

Relative importance of predictor variables:

The relative importance of predictor variables quantifies the relative contributions of each covariate to the BRT model. Relative importance is defined as the number of times a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees[107]. These contributions are scaled to sum to 100 where a higher number indicates a greater effect on the response.

C.2.4: Evaluating the BRT model predictive performance

To evaluate the BRT model predictive performance we used the following statistics[115,116]: (i) Sensitivity: a value between 0 and 1, the proportion of presences correctly identified, (ii) Specificity: a value between 0 and 1, the proportion of absences correctly identified, (iii) proportion correctly classified (PCC): a value between 0 and 1 giving the proportion of presences and absences correctly classified, (iv) Youdens J or the True Skill Statistic (TSS)[116,117]: A value between -1 and 1, with 0 indicating no skill, defined as the sum of sensitivity and specificity minus one (v) Cohen's Kappa[118]: A value between -1 and 1 measuring the proportion of agreement (0 indicating no agreement) of predicted versus observed presence and absence samples, calculated from an error matrix that cross references the number of observed and the number of predicted pixels categorised as present or absent[119], and (vi) Area under the receiver operator curve[119] (AUC): The area under a plot of the true positive rate vs. false positive rate, reflecting the ability to discriminate between presence and absence. An AUC value of 0.5 indicates random discrimination and a value of 1 indicates perfect discrimination.

To calculate statistics (i-v) it was necessary to translate the BRT logistic regression probability into a binary (0/1) classification. A threshold probability was chosen such that the model sensitivity equaled model specificity[120]. In other words we find the threshold where the

positive observations are as likely to be wrong as negative observations. It should be noted here that the choice of loss function could impact the cut off value.

All prediction statistics (i-v) were evaluated on the final optimal BRT model (see S3.2.3) for each of the 10 individual cross validation testing folds, and then averaged across all folds.

When evaluating the BRT model prediction statistics we accounted for spatial sorting bias[121], which occurs when the distance between training-presence and testing-presence locations is smaller than the distance between training-presence and testing-absence locations[121]. Not accounting for spatial sorting bias causes the predictive performance statistics to artifactually improve as μ increases and absence sites occur further away from presence sites. To remove this bias we use pairwise distance sampling[121,122] where each testing-presence site was paired with the testing-absence site that had the most similar distance to its nearest training-presence site. This procedure ensured that the training model predictive performance was evaluated on testing data that were free from spatial sorting bias. We note that removing spatial sorting bias prediction reduces estimated performance. This reflects more accurate prediction metrics, not poorly fitted models.

C.3: Pseudo-data generation

While there is no consensus on which pseudo-absence generation method best predicts true species or disease distributions, four factors are believed to have the greatest effect on the predicted distribution and thus cause bias. These are (i) the geographical extent over which pseudo-absences are generated[110,111,123], (ii) the ratio of pseudo-absences to presences[124-128], (iii) pseudo-absence contamination with true but unobserved presences[129,130], and (iv) sampling bias in presence data[95,124,131].

C.3.1: Geographical extent

Variation in the geographical extents in which pseudo-absences are generated can have a large effect on predictions and performance of distribution models. Pseudo-absences drawn from a restricted extent can produce spurious models, as their environmental space is too similar to that of the presences[110,132]. Conversely, pseudo-absences drawn from too broad an extent may result in over-prediction where one or two environmental conditions dominate, thereby providing regional discrimination but failing to identify habitat suitability at a finer scale[110,132]. Two approaches to address this issue are: (i) selecting random absences from a restricted environmental space that has been determined as unfavourable for disease transmission (using profile techniques like those previously discussed)[123,126,133], and (ii) restricting random absence points to a maximum distance from any of the presence points[110,134-136]. We chose a pseudo-absence generating scheme that partially utilises both of these approaches and restrict the generation of absences to a maximum distance (μ) from any presence point. Additionally, we generate absences based on national and sub-national evidence consensus on dengue presence or absence. Specifically, we use evidence consensus percentage values which range from -100 to 100, where -100 is a complete consensus on absence, and 100 is a complete consensus on dengue presence[2]. Using these values, pseudo-absences are generated with a density inversely proportional to the evidence consensus value at a given location except at locations with a complete consensus of dengue presence. This approach has two advantages to using environmentally restricted space (i.e., estimated with profile methods): first, it incorporates independent evidence-based knowledge on the distribution of dengue, and second, it avoids bias resulting from estimated transmission extents based on insufficiently sampled presence data.

C.3.2: Ratio of pseudo-absences to presences

The number of pseudo-absences has previously been shown to have a great effect on model accuracy[110,124,131]. Most previous research on how to distribute pseudo-absences has a consensus on using proportionally more pseudo-absences to presences[124,126], but there is a balance between too many and too few pseudo-absences. Inclusion of too few pseudo-absences leads to poorly defined areas of absence, which in turn leads to over-prediction of the disease distribution[131]. Conversely, the inclusion of too many pseudo-absences can lean to an under-prediction of the disease distribution[124].

C.3.3: Contamination bias

If a disease is known to be rare, then the generated pseudo-absence data will resemble true absences and the BRT model will be close to the true model. However, for more widespread diseases (such as dengue) there is likely to be a "contamination" bias[137,138], where some pseudo-absence points actually represent true but unobserved presences[129]. In other words, in regions where there is only a weak consensus on dengue absence, we may expect to observe presences, but currently do not have evidence for, due to the difficulties associated with surveillance of a low prevalence disease. Not accounting for this bias leads to under-prediction in locations with higher true probabilities of presence. Two methods have been developed to impute true but unobserved presences from pseudo-absence data: (i) an expectation maximisation algorithm[129,139] and (ii) model fitting using a scaled binomial loss function. Both of these methods, however, require prior knowledge of the population-wide ratio of absences to presences and assume that this value is spatially constant. These assumptions are invalid in our case where the ratio is both unknown and expected to vary spatially. To solve this problem we generated random pseudo-presence data in addition to pseudo-absence data. The pseudo-presence data were generated in the same manner as the

pseudo-absence data, except no pseudo-presences were allowed in areas lacking comprehensive evidence on the dengue presence/absence status (an evidence consensus threshold of -25 – see Supplementary Information A.4).

C.3.4: Sampling Bias

Observed presences represent the distribution of reported transmission identification rather than the actual distribution of the disease, particularly at the global scale. The BRT model does not account for the geographical distribution of presence data which can cause environmental bias in the data[95]. If this bias is not accounted for, the spatial distribution of fitted BRT model output will tend to mirror the distribution of survey efforts rather than the true distribution of the disease. Likewise, such bias will lead to a model interpretation which emphasises the importance of environmental factors in sampled areas, rather than those underlying the true disease distribution[95,124]. Our data was collected from a wide variety of independent surveys and therefore is unlikely to have a systematic sampling bias.

C.3.5: Pseudo-data generation process

From the specifications outlined above we used the following procedure to generate the pseudo-data:

Algorithm A1: Pseudo-data generation

STEP 1: The national and sub-national evidence consensus values [2] were converted to raster format and standardised (see S2.2), providing a consensus on dengue absence to presence on a scale $\{-100, \ldots, 100\}$ for each 5km × 5km pixel in a global grid.

STEP 2: A random point was created on land and restricted to a maximum distance μ from any observed presence point.

STEP 3: A uniform random variable u was generated on scale $-100 < s \leq 100$ and the point in STEP2 accepted as a pseudo-absence if $u > s$ and as a pseudo-presence if $u < s$ and $s > -25$. These conditions ensured pseudo-absences could be generated in all but complete evidence-consensus countries and pseudo-presences could only be generated in areas of dengue presence, or uncertain dengue status countries and that both are weighted by country certainty on dengue status.

STEP 4: STEP 2 and **STEP 3** were repeated to generate n_p pseudo-presence points and n_a pseudo-absence points at a distance μ.

C.4: Ensemble analysis

We have identified the number (n_a and n_p) and geographical extent (μ) of pseudo-absences and pseudo-presences as the main factors affecting BRT model prediction and have presented methods to generate pseudo-data based on these parameters in an unbiased manner. However, there is no definitive procedure to choose the optimal values of these parameters to generate the most accurate predictive map. Several studies have attempted to outline recommendations for parameterising n_p, n_a and μ[110,111,124,131,140] but none of these recommendations generalise or provide an unambiguous parameter selection strategy.

To explore the effect of changing these parameters, a sensitivity analysis was performed using different combinations of pseudo-data generating parameters n_p, n_a and μ. For the sensitivity analysis, the number of pseudo-absences (n_a) and pseudo-presences (n_p) were defined as a proportion of the total number of actual data points (8,309). The proportions used for generating pseudo-absences were 1:1, 2:1, 4:1, 6:1, 8:1, 10:1 and 12:1 and pseudo-presences were 0:1, 0.01:1, 0.025:1, 0.05:1, 0.075:1, 0.1:1. The pseudo-data were also generated within a restricted maximum distance (μ) from any actual presence point, and μ

was varied through 8 distances: 5 (~555km), 10, 15, 20, 25, 30, 35 and 40 arc degrees. All combinations of these parameter values resulted in a total of 336 ($7n_a$ x $6n_p$ x 8μ) individual input data sets and BRT models.

From the sensitivity analysis it was clear that for all parameter combinations of n_p, n_a and μ, there was a similar degree of predictive performance, with all models showing a good predictive capacity. However, this is potentially misleading, as models with similar predictive accuracy do not necessarily translate to similar predictive distributions. There can be a high variance in the predictive distribution of two models, which share a similar predictive accuracy[128,141,142]. The reason for this discrepancy is that each individual parameter combination, and resulting input data, contains some independent information about the true distribution. It follows that each parameter combination data set is a sample of the possible states of the real data distribution[143], so that all parameter combinations represent a null distribution of possible states.

Therefore, rather than selecting a single BRT model from the sensitivity analysis, we used all 336 BRT sensitivity models in an ensemble[141,142,144] and evaluate the central tendency as the mean across all 336 BRT maps.

In addition to the predictive map, all prediction (C.2.4) and summary statistics (C.2.3) were also averaged across all 336 BRT sensitivity models.

C.5: Overview of Map Generation

The fitted BRT ensemble map was produced at a 5km × 5km resolution. On this predicted map we created risk exclusion masks based on (i) the temperature suitability model

(Supplementary Information B.3.2) and (ii) the definitive extent of dengue virus transmission[2]. Pixels in which the temperature regime provided no window within an average year for completion of the extrinsic incubation period were considered at zero risk. Areas with an evidence consensus on dengue absence (<-25) [2] were masked in the final risk map.

C.6: Output maps and partial dependence plots

The final BRT map at 5km × 5 km resolution with the overlaid exclusion masks is shown in Figure SC1 below. Our map predicts a ubiquitous probability of occurrence throughout the tropics, with the highest risk in the Americas and Asia. Predicted probability of occurrence in Africa, while more unevenly dispersed than in other tropical endemic regions, is much more widespread than suggested by previous maps. Across all 336 BRT models, the average prediction performance was high (Table T2), indicating good model fits. Examination of the partial dependence curves (Figure SC2) reveals that the main predictors contributing to the occurrence map were precipitation (accounting for 26.1% of the variation explained by the model), temperature suitability (21.7%) and urban covariates (16.1% and 13.2% for the categorical urban and peri-urban demarcation respectively and 13.5% for urban accesibility). The maximum precipitation covariate caused an increase in response (probability of occurrence) up to rainfall values of around 600mm per year, after which, there is no further effect on the response. Probability of occurrence increased approximately linearly with temperature suitability. For urban accessibility there was a sharp decline in response as the travel time to a city of 50,000 persons increased, with travel times greater than 5 hours causing no effect on the response. Relative poverty (4.3%) caused a decrease in response as the GDP adjusted for purchasing power parity increased. Minimum precipitation (3.4%) and NDVI (1.65%) did not contribute greatly to the model.

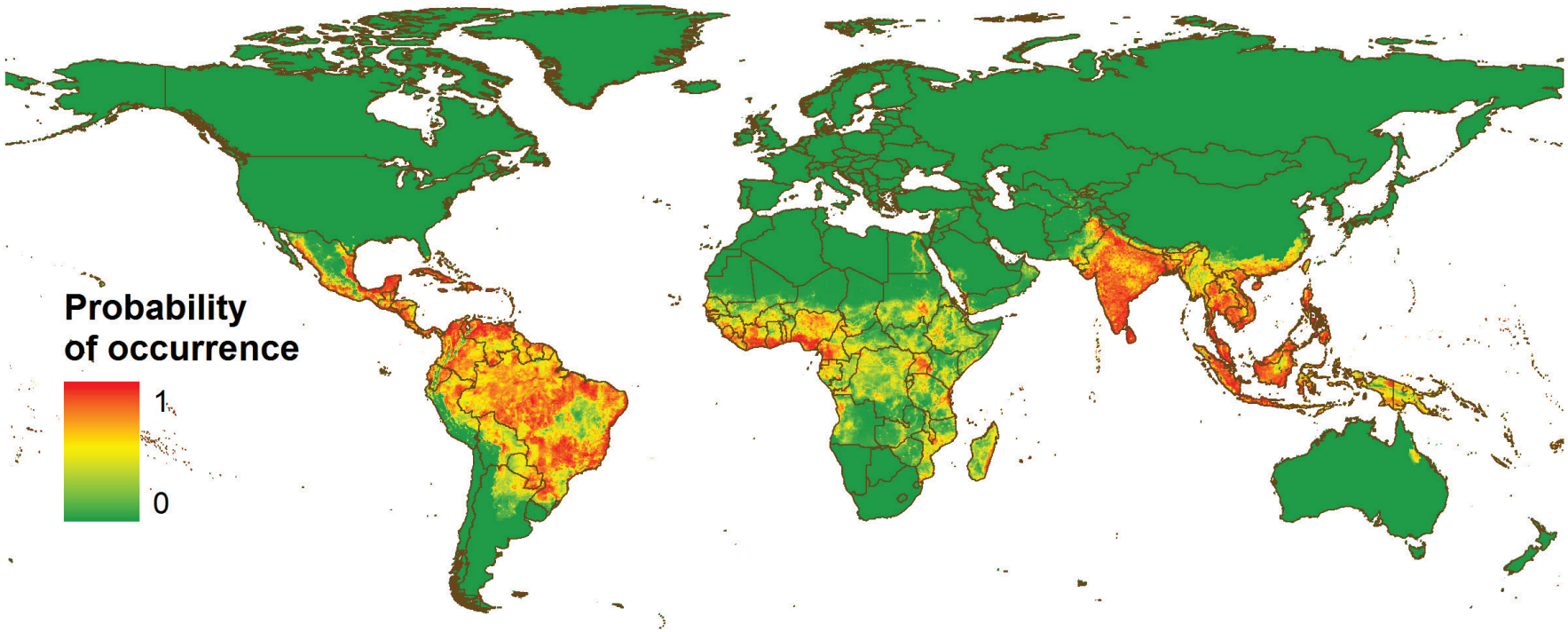

Figure SC1.BRT probability of occurrence map. Map predicted at a 5km × 5 km resolution with exclusion criteria defined in Supplementary Information C.5.

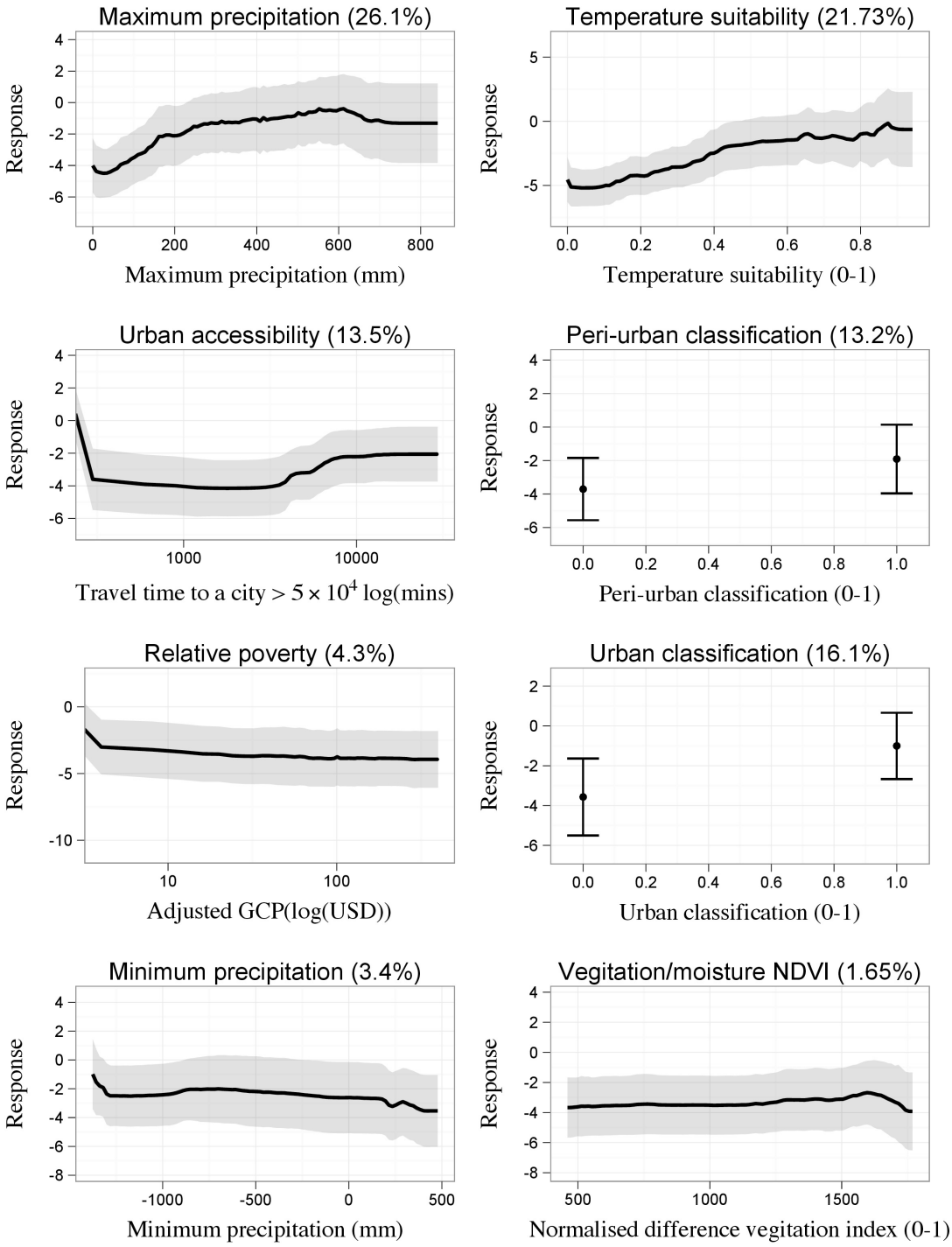

Figure SC2. Partial dependence plots averaged over all 160 BRT ensembles. Black lines represent the mean partial dependence over all 336 BRT ensembles and grey envelopes the standard deviation from the mean. The *y*-axis is the untransformed logit response and *x*-axis is the full range of covariate values. The percentage values in parentheses show the relative contributions averaged over all 336 BRT ensembles.

Table T2. **BRT prediction statistics**

	Mean	Standard Deviation
Kappa	0.51	0.036
AUC	0.81	0.020
True skill statistic	0.50	0.036
Sensitivity	0.69	0.036
Specificity	0.81	0.026
Percent correctly classified	0.75	0.017

D: Global burden and population-at-risk estimation

D.1: Overview

Despite the widely quoted figure of 50-100 million dengue infections per year[145-147], contemporary estimates of the annual global incidence of dengue have a limited evidence base. The first estimate in 1988 suggested an approximate figure of 100 million infections per year, based on assuming a ten percent annual infection rate amongst a population-at-risk of one billion[148] (Figure SD1). This annual infection rate was based on data from a small number of epidemics in the latter half of the 20th century. Although this was only ever intended as an approximate estimate, the figure of 100 million is still widely cited, despite the realisation of a much larger population-at-risk and a more variable infection rate than was originally assumed.

A revision was made in 1994 using the same methodology[149], when it became clear that dengue was far more widespread and there was increasing uncertainty over the proportion of inapparent infections. The evidence base for the four percent annual infection rate in this work is unclear and results in a lower estimated burden of 80 million infections per year. As more information became available concerning (i) the ratio of dengue haemorrhagic fever (DHF) cases to dengue fever (DF) cases and (ii) the ratio of deaths to DHF cases, a figure of 50-100 million infections globally gained more support[150,151] (Figure SD1). This figure has since been adopted by the WHO and been their estimate for the last 15 years.

In the absence of accurate or suitable estimates for the apparent-to-inapparent infection ratio and with a decline in global dengue reporting[152], progress on global burden estimation was hindered. Attention moved to estimating numbers of DHF and DF clinical cases using

generic methods of "expanding/inflating" reported DF or DHF cases. The DHF case reporting in South East Asia (SEA) was considered to be the most accurate in the early 2000s, and therefore sex-specific DHF incidence values were estimated for SEA and then extrapolated to give a global estimate of 0.4-0.5 million cases of DHF a year[153]. Using this figure and several estimates of the ratio of DHF to clinical DF cases, a global figure of 8 million cases of clinical DF was then suggested[154]. A similar figure of 9 million clinical DF cases was presented in the WHO Global Burden of Disease 2004 update, which used reported dengue deaths and separate estimates of the DHF-to-clinical DF cases ratio for SEA and the Americas[155]. Methods similar to these have also been widely used to estimate global deaths due to dengue.

While the apparent-inapparent link may, until now, have been insufficiently evidence-based, many attempts have been made to calculate national incidence estimates for the purpose of economic burden estimation[156-164]. These vary in their thoroughness, but the best approach has been to use locally relevant cohort studies to derive sex- and age-specific estimates for the apparent-inapparent ratio and then apply these to multi-year reported clinical DF datasets. A more common approach is to gather a range of these ratios, which range from 1:0.3 to over 1:100, and then apply them to national clinical dengue case data of variable reliability. The major problem with this approach is geographical variation in dengue transmission intensity, treatment-seeking behaviour and healthcare treatment and reporting capacity. A factor that is often overlooked is that many of the cohort studies employ active fever surveillance in their study population which reduces barriers to healthcare access, thus modifying treatment-seeking behaviour[165]. This ensures simply converting nationally reported numbers of clinical dengue cases to infections using a common factor is likely to give a significant underestimate of true infection incidence.

Although the inflation factor approach is unsuitable on a global scale, this method of inference based upon cohort studies has reinforced the estimate of 50-100 million infections. Of the 2.2 million clinical dengue cases reported to the WHO in 2010, 1.7 million of these were reported in the Americas[166]. Therefore, using an average range of previously-used expansion factors for this region (6-27)[159,164,167] would suggest that between 10 and 46 million dengue infections occur in the Americas alone[164]. The variability in national clinical dengue case data has prompted the latest estimate, which excludes it altogether. Beatty *et al.*[168] redefined a dengue-endemic country as having a published record of dengue occurrence or sharing a significant border with one that does. This suggested a population-at- risk of 3.5 billion from which average inflation factors suggest a total of 100-200 million dengue infections and 36 million cases of clinical DF per year[168,169]. This generalized single factor global approach is unlikely to give accurate national case burden estimates, however, their focus on creating a solid evidence base for the global extent of the disease creates a better basis for initial burden estimates than approaches that use reported case data of variable or unknown quality.

Considering the variable data sources on which these estimates are based, it is perhaps surprising that confidence intervals are not presented. With the exception of Rigau-Perez *et al.*[150] and Beatty *et al.*[168,169], no existing estimates have conveyed the uncertainty present in dengue burden estimation (Figure SD1), yet we consider this a vital step for accurate interpretation of these estimates.

Our study has produced the first cartographic based burden estimate for dengue. We assembled an extensive data set of 54 cohort studies defining incidence rate in person years

with estimates of the inapparent-to-apparent ratio. Using a hierarchical Bayesian model, we estimated a relationship between the cohort incidence data and our previously generated BRT (boosted regression trees see Supplementary Information C) map of the probability of occurrence of dengue infection (see main text), following a rubric applied previously to malaria[170,171]. Using this relationship, we provide estimates of annual inapparent and apparent dengue infections with confidence intervals at the national, continental and global scales.

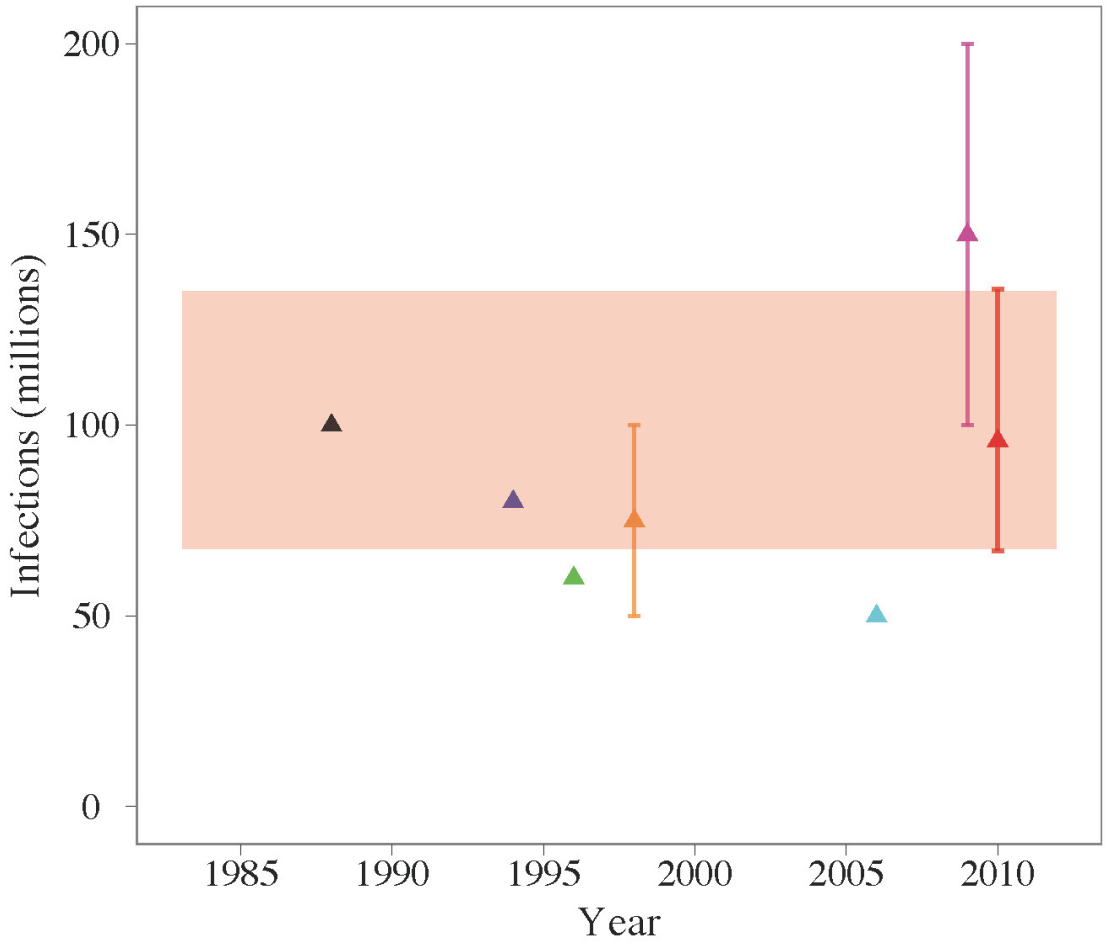

Figure SD1. Global estimates of dengue infections. Comparison of previous estimates of total global dengue infections in individuals of all ages, 1985 to 2010: ▲ Halstead *et al.* 1988[148], ▲ Monath *et al.* 1994[172], ▲ Rodhain *et al.* 1996[173], ▲ Rigau-Perez *et al.* 1998[150], ▲ TDR/WHO. scientific working group 2006[174], ▲ Beatty *et al.* 2009[175], ▲ apparent infections from this study. Estimates are aligned to the year of estimate and, if not stated,

aligned to the publication date. Red shading marks the credible interval of our current estimate for comparison.

D.2: Assembly of cohort studies

D.2.1: Existing incidence data

Estimating incidence of dengue fever is complicated by a spectrum of clinical manifestations[176] and variable reporting capacity of different healthcare systems. Therefore, the most accurate way of comparing incidence rates in different locations worldwide is through serological cohort studies following a standard methodology. Due to the high proportion of inapparent dengue infections[177], active case detection must go beyond improved clinical surveillance. Estimates of total cohort dengue infections must observe immune responses to dengue virus antigens before and after the dengue transmission season [178]. Dengue virus-infected humans exhibit dengue-specific Immunoglobulin M (IgM) and Immunoglobulin G (IgG) immune responses. High IgM titres can be observed in primary infections as soon as four days[179] after the onset of fever and persist for another 30-90 days[180]. IgM responses can also be seen in secondary dengue infections, but the response is often slower, weaker and shorter lived[181]. In contrast IgG levels rise around seven days[179] after the onset of fever in primary infection and can be observed for life[181] thus providing a good indicator of previous dengue exposure. The total number of primary infections can be obtained by observing the number of IgG-negative individuals that seroconvert to IgG-positive status before and after the transmission season. The number of post primary dengue infections can be estimated by identifying individuals with both IgG and IgM responses after the dengue transmission season. This is likely to under-estimate post primary infections due to weaker secondary IgM responses[181] and the 30-90 day window for IgM detection[180]. Experimental protocols often assume infection is the result of certain predominant serotypes

that may be circulating in a given study region at the time and thus monitoring may be type-specific, resulting in further underestimation if multiple serotypes are circulating. IgM and IgG serologic surveys also cross-react with other flaviviruses[182] and therefore need situation-specific controls to estimate sero-conversions to dengue alone. However, despite these limitations, serological cohort studies represent the best possible estimation of local incidence for dengue.

Cohort studies are often difficult to compare due to varying demographic dynamics. For dengue, while higher incidence has been reported in paediatric populations[183], adult populations have exhibited higher incidence in other settings, so the demographic distribution of dengue infection remains unclear. Therefore, having no robust function with which to adjust for age and no clear demographic basis for exclusion of available information, our assemblage of cohort studies comprised incidence in all age groups. The inclusion criteria developed for the cohort study database are described below.

D.2.2: Inclusion criteria

(i) The study took place during or after 1960 to coincide with the date range of our occurrence database.

(ii) Surveys were longitudinal and involved active case detection of sero-conversion to dengue type-specific antibodies in a defined cohort.

(iii) Monitoring of sero-conversion through paired blood samples was undertaken at least before and after each dengue transmission season for IgG immune responses, or at least every 90 days for IgM immune responses[180,184].

(iv) Data was presented in a way that enabled the total number of infections and the number of person-years of observation to be obtained.

(v) All dengue infections were identified from blood samples that distinguished primary and secondary infections using all or one of the following methods: (a) Hemagglutination inhibition (HI), (b) plaque reduction neutralization test (PRNT) or (c) enzyme-linked immunoabsorbant assay (ELISA).

(vi) Surveys were conducted over at least a 12 month period, or else over a clinically defined period of transmission with blood samples taken before and after this period.

The location, total number of infections, person-years of observation and the ratio of inapparent to apparent (I:A) infections was recorded. Definitions and methods for detecting apparent infections varied from study to study but adhered to the following general criteria: (i) an apparent case was any manifestation of febrile illness accompanied by fever greater than 38^{o}C; (ii) such infections were detected either through enhanced clinical surveillance, retrospective cohort participant questionnaires, or systematic surveillance of school or workplace absentees. Therefore our definition of an inapparent infection is an infection that does not have any impact on the day-to-day life of the subject. As such, an inapparent infection will not modify a person's regular schedule e.g. attending school, register as an extraordinary period of ill-health that can be recalled when later questioned, nor will it prompt any treatment-seeking beyond self medication. While each of the separate symptom detection methods has the potential to underestimate apparent infections, they allow us the best possible estimate of the I:A ratio.

D.2.3: Summary

A search for "dengue cohort study" in PubMed, subsequent reference tracking, and personal requests enabled the identification of 55 geographically unique locations from 38 cohort studies in 19 countries in a variety of regions. Estimates of the I:A ratio were available for 40

locations (27 cohort studies, 15 countries). We excluded one cohort study by Teixeira *et al*[185], as both the reported incidence and I:A ratio were an order of magnitude outside the others observed. Excluding this study, the mean reported incidence across all cohort studies was calculated as 129.7 per 1000 person years (standard deviation ±135) and the mean I:S ratio as 4.3 (standard deviation ± 2.8). The incidence and I:S ratio from Teixeira *et al*[185] was 706 and 1555.55 respectively, which we deemed implausible given the other cohort studies.

D.3: Relationship between incidence and probability of occurrence

We determine the relationship between incidence and probability of occurrence using a Hierarchical Bayesian linear model[186]. We chose a Bayesian formulation due to statistical robustness, explicit handling of uncertainty, transparent variable and model selection and the ease of incorporating complex nonlinear functions[186]. The Bayesian hierarchical model is defined as a tiered structure where, at the first level, a likelihood function defines the probability distribution that generates the data (the data model - *[data|process,prameters]*) , at the second level, prior distributions define the parameters of the likelihood function (the process model - *[process|parameters]*), and at the third, and final level, hyper prior distributions define the prior parameters (the parameter model - *[parameters]*). The end result of the product of these three distributions is proportional to the posterior distribution which is the distribution of the process and the parameters *[process,parameters|data]*.

D.3.1: Data model

Modelling count data (incidence per 1000 person years) imposes restrictions on the choice of probability distribution as an event count is the realization of a nonnegative integer-valued random variable[187,188]. The foundation building block in this modelling framework is the

Poisson regression model where the variance of a random variable is constrained to equal the mean[188]. However, a more broadly applicable and general specification – the negative binomial model[170,189] can be used to include the case when the variance exceeds the mean. We choose this distribution to ensure maximum flexibility in modelling our data:

$$p(x|f(x),r(x)) = \frac{\Gamma(x+V)}{x!\,\Gamma(V)}\left(\frac{V}{f(x)+V}\right)^{V}\left(\frac{f(x)}{f(x)+V}\right)^{x} \quad (2)$$

where x is the incidence per 1000 person years, $f(x)$ is the mean or rate function defining how the mean value of incidence changes with the probability of occurrence and $r(x)$ is the noise or dispersion function, representing the variance in population-wide levels of incidence.

D.3.2: Process model

We model the negative binomial rate function, $f(x)$, using a Gaussian process[170,190]. The Gaussian process model ensured that no parametric form was imposed on $f(x)$, thereby allowing for a data-driven approach. In this context, the Gaussian process was parameterised with two components: a mean function ($M(x)$, controlling the central tendency of the function at a given value of x) and a covariance function ($C(x,x')$), controlling the second order characteristics of the function, such as differentiability). For M a quadratic function: $M(x) = Ax + Bx^2$ was used and for C we chose a highly smooth Gaussian kernel[191] characterised by two parameters that control for the scale (*Scale*), and amplitude (*Amp*) of the covariance[190].

We included several constraints to prevent biologically implausible scenarios from being included in the model. First the Gaussian process was constrained to include positive values only if $(f(x) > 0\ \forall\ x)$ to prevent impossible negative incidence values. Second the Gaussian process was conditioned to include the assumption that at zero probability of occurrence,

there is zero incidence $(f(0) = 0)$. Third, the Gaussian process was constrained to have at most one inflection point (excluding a saddle inflection point), thereby allowing only ecologically simple models without multiple peaks, troughs or saddles.

Previous approaches have modelled the dispersion or noise function[192] as a quadratic function incorporating specific prior biological knowledge about the disease being modelled[170]. For our dengue incidence data we lack any such prior knowledge, and therefore to prevent any bias and unjustifiable complexity we parameterise the dispersion function as constant, that is $r(x) = V$.

D.3.3: Parameter model

Hyper priors for the Gaussian process rate function parameters, $\theta_{f(x)} \in \{A, B, C, Scale, Amp\}$, were set to uninformative[186] uniform distributions with sensible ranges. This was done to express vague prior knowledge about these parameters. The hyper parameter for the dispersion function, $\theta_{r(x)} \in \{V\}$, was set to an uninformative hyper prior [186,192] requiring no defined ranges.

D.3.4: Posterior inference

The model was fitted and the posterior characterised using Markov Chain Monte Carlo sampling (MCMC)[192-194]. The final models were run over one million iterations, sampling every 500 iterations to prevent autocorrelation effects between samples[186]. Additionally, the first 200,000 samples were discarded ("burn in") to ensure that the posterior was drawn from the equilibrium distribution of the Markov chain. This gave a total of n=1600 posterior samples for all the model parameters. Visual inspection of the MCMC trace and Geweke plots[192,195] were used to check model convergence. The output of the burden model consists

of samples from the joint posterior distributions for all the model parameters $\theta_i \in \{f(x)_i, r(x)_i, A_{i,}, B_i, C_i, Scale_i, Amp_i, V_i\}$ where $i = 1,,,n$. For all individual combinations of these parameter samples (θ_i)

individual curves can be constructed, using the data model (equation 2), that represent a realised relationship between incidence and the probability of occurrence (Ω). The full set of these realisations ($\bar{\Omega} = [\Omega_1, \ldots, \Omega_n]$) represents a joint model for the data and unknown parameters.

The entire model fitting procedure was performed separately for apparent and inapparent infections, resulting in two separate burden models derived from the I:A ratio measured in the serological cohort studies. It should be noted that only 39 of the 54 cohort studies provided information on these ratios and, therefore missing ratio values were imputed in the MCMC[192]. The results of the burden models are shown in Figure SD2.

D.4 Overview of map generation and burden estimates

Each burden model (apparent and inapparent) provided a posterior set of relationships ($\bar{\Omega}$) between the probability of occurrence and incidence. From each realised relationship (Ω_i) the probability of occurrence map was translated into an incidence map. We did this for all relationships in $\bar{\Omega}$ generating a distribution of incidence maps, reflecting the posterior predicted relationships. For each of these maps, using a human population surface for 2010[80], infection numbers were estimated on a global scale. Stratification of these estimates into national and subnational divisions is discussed in Supplementary Information E.

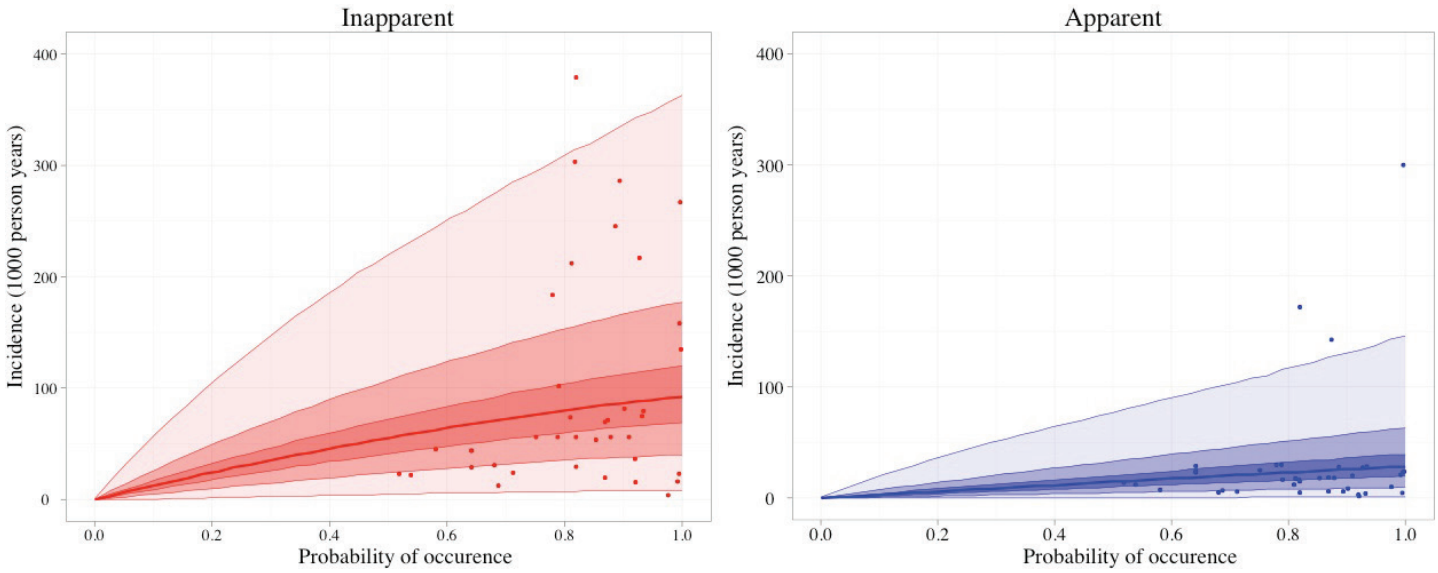

Figure SD2. Bayesian modelled relationship between the probability of occurrence and incidence for inapparent and apparent number of infections. The data are the points, the bold lines are the medians and the envelopes are the 0.25, 0.5 and 0.95 credible intervals centered on the median displayed with progressively lighter shades.

E: Reconciling cartographic and surveillance-based burden estimates

E.1 Overview

This section provides a detailed summary and discussion of apparent and inapparent dengue infections at a country level as estimated by our cartographic methods described in Supplementary Information D. These are compared to surveillance-based burden estimates of dengue cases reported to the World Health Organization (WHO)[166,196-198]. Here we explain how the order of magnitude difference between these two estimates is feasible if we consider the systematic underreporting introduced at multiple stages along the pathway that separates an apparent infection in the community with a reported clinical dengue case at the national level. We then go on to explain and quantify each of these steps in detail using a comparative analysis of previously published estimates for each step to reconcile the two differing burden estimates. E.2 describes the data sources used in the surveillance-based method for cases reported to the WHO. E.3 describes the global distribution of apparent dengue infections as estimated by the cartographic approach and clinical cases as reported to the WHO using the surveillance-based approach. In E.4 the relevant data loss steps are discussed to reconcile these two estimates.

E.2 Surveillance-based burden data sources

National annually reported numbers of diagnosed dengue cases were taken from the WHO regional office websites[166,196-198]. While DengueNet[199] remains the central repository for dengue data within the WHO, its accessible database contains only sporadic case numbers since 2005. By obtaining case numbers directly from WHO regional offices, we were able to

obtain a more contemporary estimate of national burden. For this analysis we used the average annual case numbers from the three most recent complete years available — 2009-2011 for Pan American Health Organization (PAHO) and Eastern Mediterranean Region Office (EMRO), 2008-2010 for South East Asian Region Office (SEARO) and Western Pacific Region Office (WPRO). Case numbers presented by PAHO (47 countries) and SEARO (11 countries) are based exclusively on country reports to these offices, which are themselves based around the WHO clinical guidelines[176]. WPRO case numbers come from country reports to WPRO (23 countries) and country Ministry of Health websites (4 countries). Case numbers reported by EMRO (1 country) are from the WHO country offices and are only available in sporadic reports. Imported cases were excluded when a differentiation between imported and indigenous cases was made. There is no legal obligation for countries to report annual case numbers to the WHO, nor is any attempt made to standardise what countries report beyond issuing standard guidelines[176]. While diagnostic limitations and a lack of standardisation make international comparisons difficult for dengue, we believe that comparing country reports of diagnoses based on WHO guidelines forms the most reliable and internationally comparable source for annually reported diagnosed dengue cases.

Here we are interested in comparing surveillance-based estimates of total symptomatic clinical dengue cases (dengue fever (DF) + dengue hemorrhagic fever (DHF) + dengue shock syndrome (DSS) or dengue (DW^-) + dengue with warning signs (DW^+) + severe dengue(SD)) with our own cartographic burden estimates of total apparent infections [176,200]. While countries are instructed to report only laboratory confirmed cases large case numbers often make it prohibitively expensive to laboratory-confirm every suspected clinical case. It is therefore not uncommon for countries to report suspected dengue (based on clinical

diagnosis) and confirmed dengue (based on laboratory diagnosis) cases as a single combined figure. The distinction into DF+DHF+DSS or DW^{-}+ DW^{+}+SD is then often made using a country-specific adapted WHO case definition which may or may not be reported in the WHO figures (Table T3) [201,202]. In this analysis we took the broadest spectrum of clinical manifestations available in the WHO data, however it is well known that reporting fidelity and standardisation of the clinical forms is far from globally consistent and our comparison with cartographic burden estimates must take these regional differences in surveillance into account (Table T3).

Table T3. Levels of reporting in the four World Health Organization (WHO) regions that display dengue data. PAHO = Pan American Health Organization, WPRO = Western Pacific Region Office, SEARO = South East Asian Region Office, EMRO = Eastern Mediterranean Region Office, DF = dengue fever, D = dengue, SD = severe dengue.

WHO region	Total DF/D	DF(suspected/confirmed)	SD	Deaths	Percentage of global dengue cases reported 2008-2010
PAHO	✓	✓	✓	✓	73
WPRO	✓	✓		✓	14
SEARO	✓	✓*		✓	12
EMRO	✓			✓	1

* Available for some countries for some years.

E.3 Country-level burden estimates

In 2010, the total global apparent dengue infection burden estimated in this study (96 million, credible interval = 67-136) is substantially larger than the global number of reported clinical dengue cases (2.2 million). However, the burden rank of each country is largely consistent in both estimates and the differences in absolute burden estimates often show a common factor that suggests an intrinsic loss to underreporting.

In the countries of the Americas, both approaches predict a concentration of the burden in four countries: Venezuela, Colombia, Mexico and Brazil (Figure SE2). However, in contrast to the WHO figures, our estimate suggests an additional sizable contribution from Peru, Guatemala and the Caribbean islands of Cuba and the Dominican Republic. Outside of these countries both approaches agree on a uniformly low burden. Our estimate for total apparent infections in the Americas (13.3 million) is around seven times that reported by the WHO.

WHO estimates were not available for the African regions, with the exception of Cape Verde. For Cape Verde, our estimate of 10,515 cases (Figure SE3) is comparable with the WHO reported figure of 10,760. Elsewhere, we predict the highest burden in the areas of high population density in the countries of Nigeria, Egypt, Democratic Republic of the Congo, Ghana and Uganda. In addition, total apparent burden is considerable in a large number of countries, with 32 having a burden of over 50,000 apparent infections per year, contributing an additional 15.7 million apparent infections.

In Asia, both approaches predict a high burden in South East Asia and the Indian subcontinent (Figure SE4). Our estimates from countries in South East Asia are consistently around 40 (20-60) times those reported to the WHO. Despite a large number of countries falling within this interval, our figures for China (6.5 million) and India (32.5 million) are far above what is reported to the WHO. In both cases, large populations are combined with large areas of high suitability for dengue. We suggest that under-reporting of dengue in these two countries is a significant part of reconciling the gap between our global burden estimates and cases reported to the WHO. China and India contribute 58% of the total burden for Asia (66.8 million apparent infections per year). It is clear that reducing the uncertainty in the estimates

for these two countries is important for reducing uncertainty for total global estimates. Achieving this will require better evidence consensus in low consensus areas, more occurrence points, particularly where our data is currently sparse, and additional cohort studies across a range of transmission intensities.

In Oceania, both approaches predict a low burden, with our predictions suggesting that only Papua New Guinea faces a burden of over 50,000 apparent infections a year (Figure SE1). Considering their population size, Fiji and Samoa also contribute a sizable portion of Oceania's total 178,000 apparent infections. Our predictions are, on average, 40 times greater than the WHO estimates, although this is variable due to low reported case numbers.

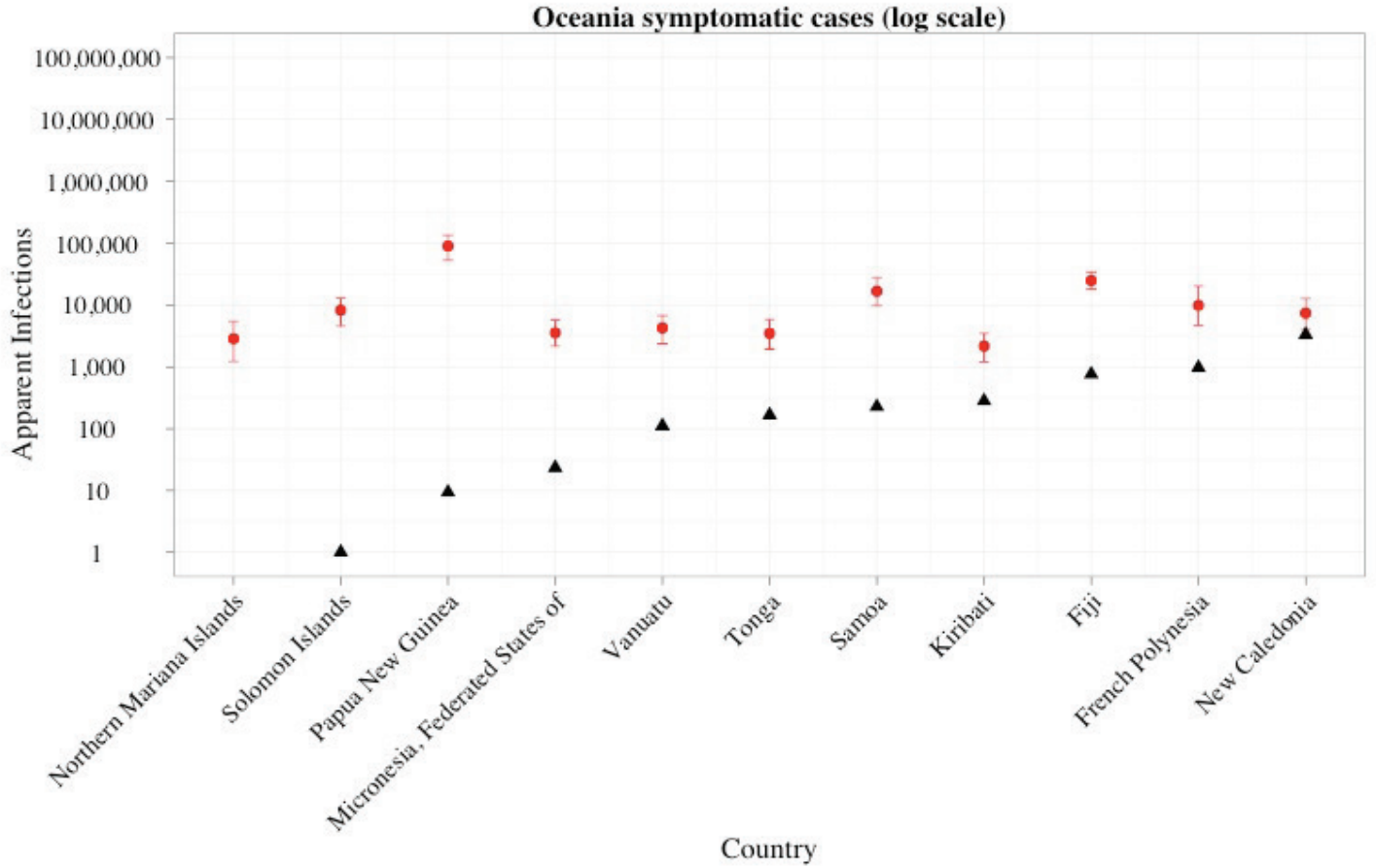

Figure SE1. Number of apparent infections (red) and number of diagnosed dengue cases reported to the WHO (black) per country in Oceania.

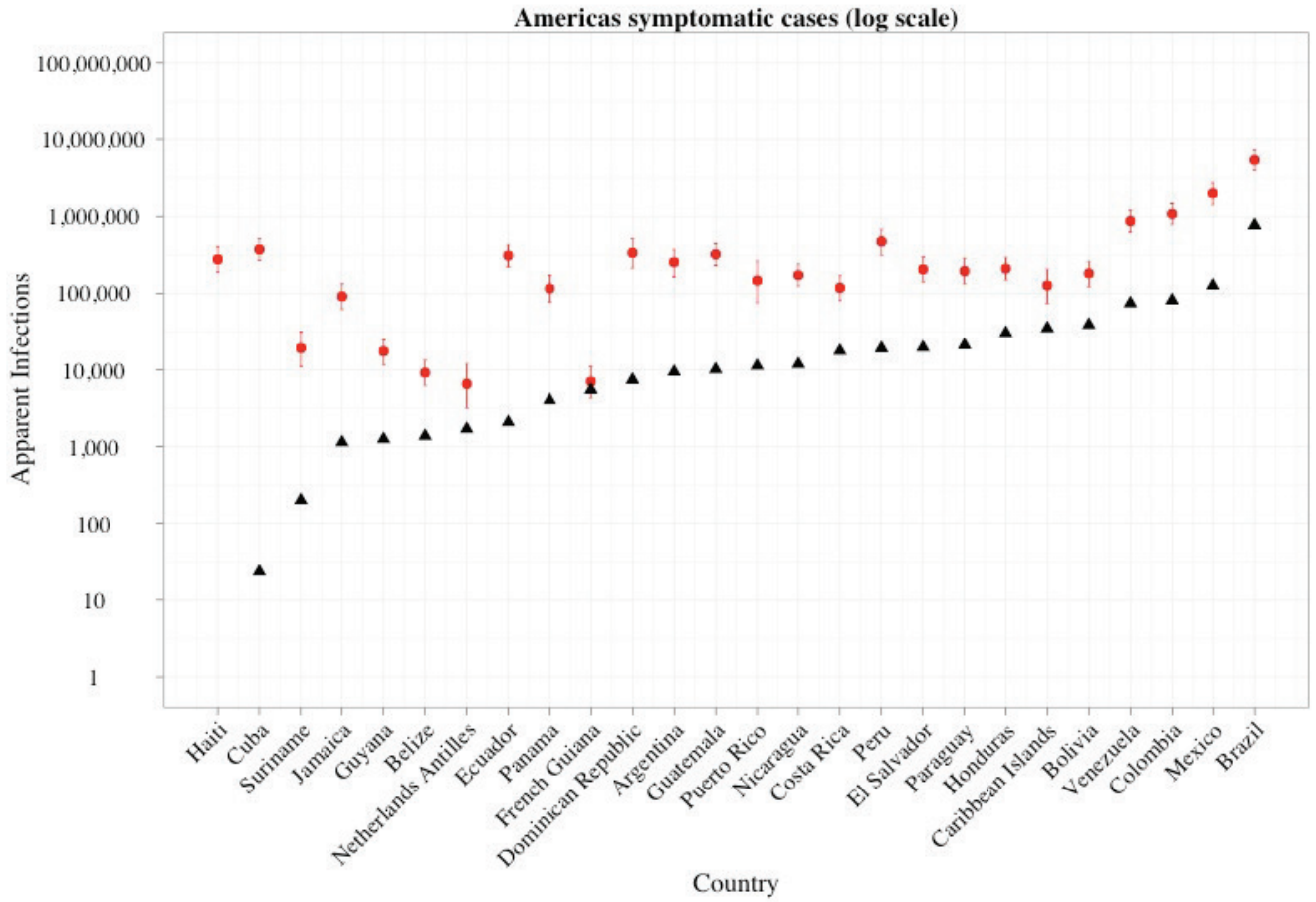

Figure SE2. Number of apparent infections (red) and number of diagnosed dengue cases reported to the WHO (black) per country in the Americas.

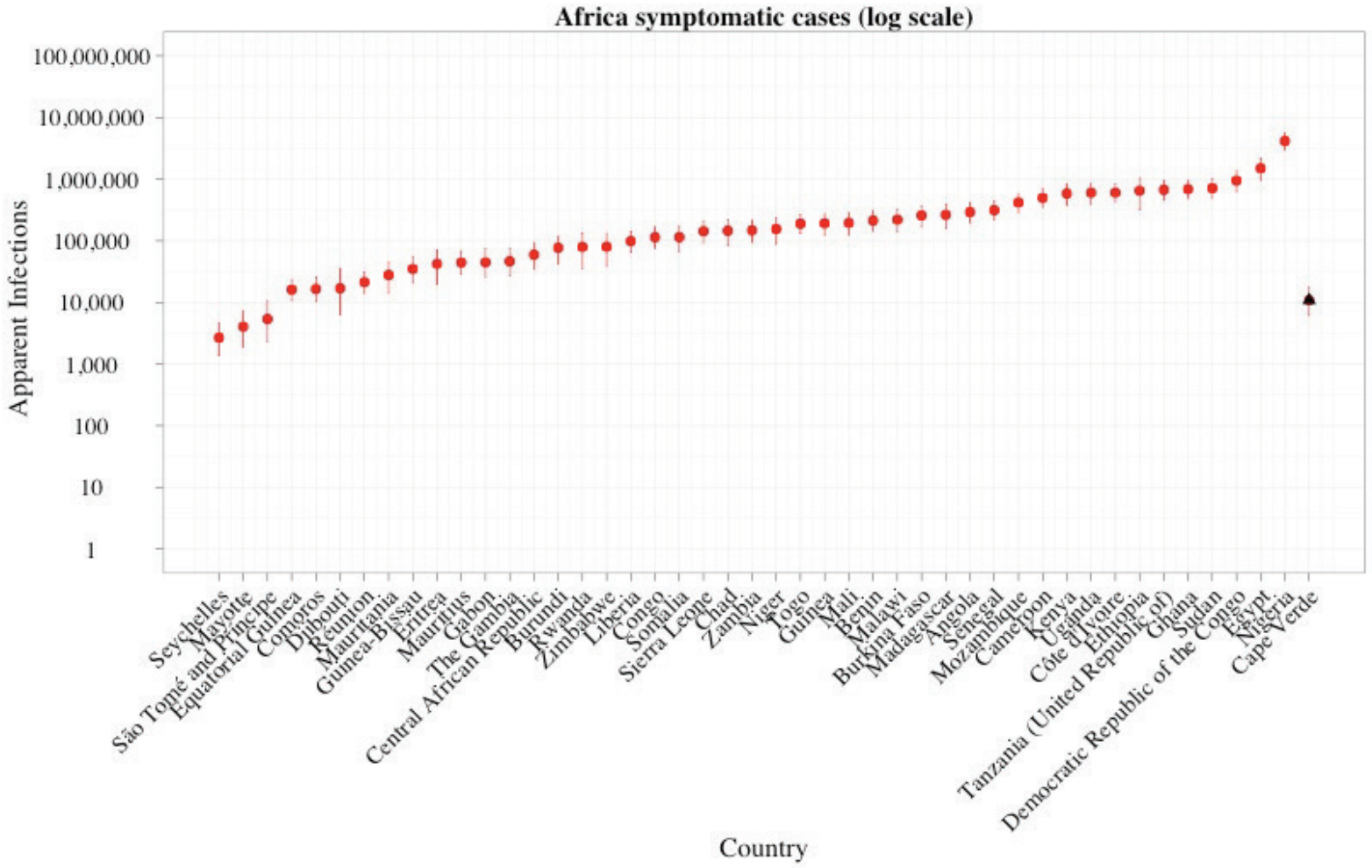

Figure SE3. Number of apparent infections (red) and number of diagnosed dengue cases reported to the WHO (black) per country in Africa. Only Cape Verde provided any reported cases to the WHO.

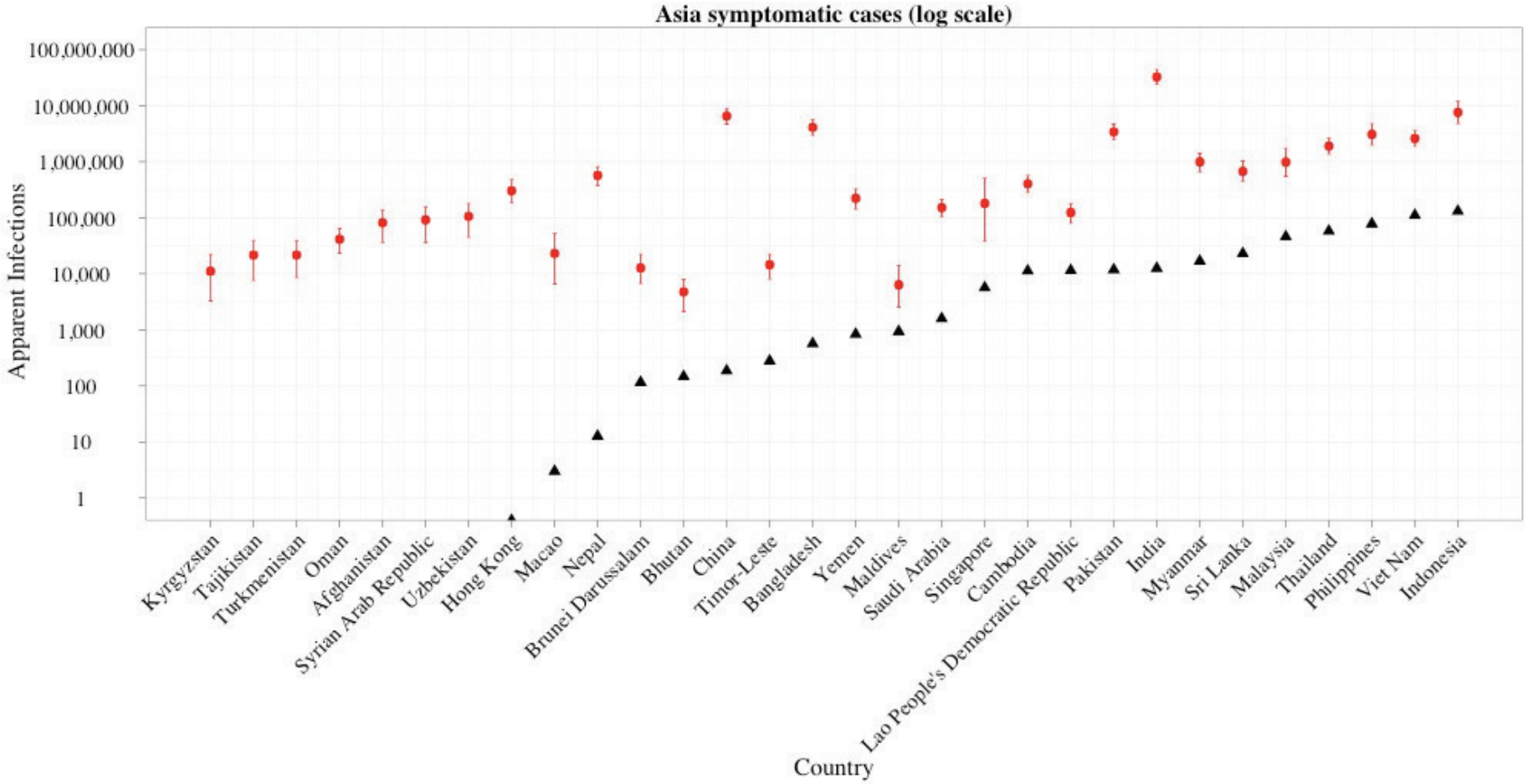

Figure SE4. Number of apparent infections (red) and number of diagnosed dengue cases reported to the WHO (black) per country in Asia.

Country Name	Apparent Mean	Apparent 2.50%	Apparent 97.50%	Inapparent Mean	Inapparent 2.50%	Inapparent 97.50%	WHO. estimates	CI Ratio Rank	CI Absolute Rank
Aruba	3,165	1,451	5,820	9,715	5,174	16,299	1,762	22	116
Afghanistan	81,687	36,803	138,574	255,681	128,224	399,333		33	63
Angola	292,178	196,712	419,141	896,115	643,138	1,210,316		107	37
Anguilla	519	216	1,006	1,571	770	2,767	1,418	16	133
Netherlands Antilles	6,546	3,199	11,930	19,765	10,998	32,919	1,690	24	104
Argentina	254,470	162,631	370,798	787,499	532,735	1,082,768	9,337	85	41
American Samoa	2,414	935	4,883	7,316	3,421	13,350	373	12	119
Antigua and Barbuda	3,226	1,592	5,800	9,739	5,467	16,031	3	28	118
Burundi	77,023	42,810	119,252	238,713	145,122	345,040		62	71
Benin	213,030	143,875	311,527	651,489	469,940	894,900		96	46
Burkina Faso	257,950	167,531	370,996	797,426	548,974	1,084,392		95	42
Bangladesh	4,097,833	2,952,879	5,608,456	12,581,091	9,519,133	16,359,636	568	135	7
Bahamas	10,971	6,306	18,003	33,335	21,183	50,405	2,336	53	101
Belize	9,128	6,208	13,270	27,878	20,264	38,061	1,368	101	108
Bolivia	181,219	122,119	260,474	555,702	399,308	751,368	38,640	106	53
Brazil	5,371,268	3,952,287	7,283,317	16,404,160	12,672,363	21,111,729	765,769	139	4
Barbados	9,398	4,753	16,653	28,343	16,213	46,077	1,239	32	100
Brunei Darussalam	12,732	6,776	22,541	38,421	22,836	62,606	116	35	94
Bhutan	4,793	2,109	7,946	15,042	7,343	23,177	147	38	111
Central African Republic	59,436	34,672	93,206	183,305	117,623	266,963		63	74
China	6,523,946	4,683,881	8,919,000	20,062,625	15,083,630	26,126,115	186	133	3
Cote d'Ivoire	603,431	427,796	845,552	1,843,448	1,383,217	2,435,397		122	27
Cameroon	497,871	350,920	701,594	1,521,543	1,132,524	2,022,270		120	29
Democratic Republic of the Congo	947,801	629,617	1,356,953	2,917,767	2,062,475	3,941,733		103	15
Congo	114,173	75,677	168,494	347,147	246,770	480,451		88	67
Cook Islands	712	225	1,766	2,155	827	4,711	60	3	129
Colombia	1,073,891	783,699	1,465,285	3,281,089	2,515,002	4,243,729	80,634	137	17
Comoros	16,538	10,228	25,414	50,733	34,230	72,452		73	96
Cape Verde	10,879	6,126	17,651	33,547	20,972	50,226	10,760	55	102
Costa Rica	117,677	79,912	168,993	361,243	261,580	488,074	17,524	109	69
Cuba	372,825	266,724	518,849	1,132,115	856,787	1,489,340	23	123	34
Cayman Islands	1,987	1,058	3,404	6,016	3,599	9,452	3	41	126
Djibouti	16,946	6,291	35,062	52,407	23,568	97,101		11	86
Dominica	2,327	1,286	3,880	7,052	4,347	10,850	227	47	125
Dominican Republic	336,410	213,296	519,207	1,017,290	698,134	1,466,630	7,383	76	30
Ecuador	310,448	220,963	430,732	951,375	713,884	1,248,114	2,066	124	40
Egypt	1,499,568	965,744	2,164,954	4,645,241	3,186,358	6,312,307		91	13
Eritrea	42,184	19,704	70,386	131,736	68,901	202,693		39	75
Ethiopia	651,184	320,718	1,041,952	2,037,422	1,089,989	3,038,520		48	16
Fiji	24,969	18,152	34,109	76,371	58,437	99,072	759	136	93
Micronesia, Federated States of	3,567	2,158	5,701	10,872	7,153	16,136	23	61	121
Gabon	44,792	25,942	74,225	135,942	86,746	208,083		51	77
Ghana	687,110	486,967	963,366	2,093,455	1,571,708	2,765,615		121	22
Guinea	192,067	125,712	275,871	593,001	412,489	803,659		97	50
Guadeloupe	17,466	10,027	28,528	52,680	33,409	79,826	14,754	56	89
The Gambia	46,476	27,000	76,385	141,685	90,603	214,567		54	76
Guinea-Bissau	35,011	20,615	55,465	107,616	69,781	157,739		60	82
Equatorial Guinea	16,166	10,729	23,256	49,799	35,252	67,404		100	99
Grenada	3,723	1,891	6,526	11,246	6,444	18,119	78	34	114
Guatemala	322,243	231,037	443,795	988,330	745,518	1,289,753	10,073	127	39
French Guiana	7,024	4,325	11,090	21,312	14,285	31,274	5,449	68	109
Guyana	17,416	11,728	24,596	53,779	38,306	71,810	1,249	114	98
Hong Kong	304,782	184,690	475,819	924,234	613,579	1,342,712	0	69	32
Honduras	209,834	149,848	291,525	641,409	483,737	841,072	30,134	125	52
Haiti	276,581	188,402	403,229	844,925	613,842	1,152,883		98	38
Indonesia	7,590,213	4,798,222	11,944,976	23,009,108	15,724,054	33,745,901	130,575	70	2
India	32,541,392	23,809,852	44,196,670	99,692,319	76,480,648	128,730,948	12,484	138	1
Jamaica	90,807	61,553	132,005	275,459	199,486	376,774	1,132	99	73
Kenya	583,960	376,348	843,317	1,807,001	1,234,459	2,461,954		92	24

Kyrgyzstan	11,135	3,300	22,437	35,093	12,827	63,199		10	88
Cambodia	404,533	282,589	568,752	1,243,325	918,191	1,649,357	11,247	119	33
Kiribati	2,173	1,215	3,475	6,712	4,170	9,958	280	57	127
Saint Kitts and Nevis	1,845	968	3,134	5,581	3,312	8,766	23	42	128
Lao People's Democratic Republic	124,006	79,970	178,093	383,905	263,083	521,000	11,431	94	64
Liberia	98,678	65,906	140,777	304,137	216,716	408,522		108	72
Saint Lucia	6,144	3,655	9,698	18,609	12,168	27,335	226	65	110
Sri Lanka	673,544	445,991	1,027,403	2,042,226	1,447,987	2,910,693	22,902	80	18
Macao	23,158	6,626	52,502	69,833	26,037	140,928	3	5	78
Madagascar	264,443	159,191	393,246	821,514	526,343	1,148,447		79	35
Maldives	6,372	2,557	13,981	19,735	9,298	38,567	933	9	103
Mexico	1,987,320	1,422,381	2,730,919	6,102,891	4,582,683	7,964,033	125,217	128	10
Marshall Islands	1,891	757	4,195	5,774	2,687	11,424		8	122
Mali	194,903	125,686	281,231	602,552	411,989	820,354		93	48
Myanmar	992,954	669,765	1,408,977	3,056,420	2,191,868	4,098,171	16,824	111	14
Northern Mariana Islands	2,862	1,235	5,493	8,704	4,406	15,167		18	117
Mozambique	418,090	287,800	586,770	1,285,737	935,627	1,707,800		118	31
Mauritania	27,859	14,145	44,508	86,922	48,185	129,140		50	85
Montserrat	178	73	342	545	269	952	1	17	137
Martinique	14,845	8,099	25,136	44,766	27,271	69,900	12,918	44	92
Mauritius	44,471	28,232	68,107	134,755	92,595	192,540		78	81
Malawi	220,050	139,738	319,955	680,097	461,604	931,479		84	45
Malaysia	983,619	546,225	1,746,771	2,969,671	1,817,360	4,835,731	45,664	37	12
Mayotte	4,049	1,886	7,286	12,445	6,734	20,417		25	112
New Caledonia	7,423	3,988	12,665	22,609	13,639	35,448	3,301	43	105
Niger	155,313	87,805	237,801	482,943	291,858	693,142		67	51
Nigeria	4,153,338	3,004,606	5,700,852	12,698,054	9,670,162	16,510,850		132	6
Nicaragua	172,439	124,002	239,068	526,486	399,249	689,249	11,763	126	60
Niue	36	17	64	108	59	178	1	23	139
Nepal	571,773	377,060	813,702	1,769,014	1,232,877	2,386,338	13	105	26
Nauru	303	62	812	915	260	2,138	1	2	134
Oman	41,524	23,179	63,745	129,341	77,707	185,832		66	80
Pakistan	3,414,749	2,455,183	4,680,780	10,481,756	7,898,303	13,642,888	11,787	131	8
Panama	115,465	77,189	171,413	349,933	250,842	488,023	3,979	87	66
Peru	472,445	316,552	677,480	1,454,164	1,038,618	1,964,217	19,005	104	28
Philippines	3,076,863	1,990,758	4,810,993	9,339,425	6,493,806	13,584,794	77,598	74	5
Palau	715	248	1,562	2,168	920	4,207	72	7	130
Papua New Guinea	89,943	53,076	134,815	279,597	175,330	394,470	9	77	70
Puerto Rico	146,564	75,342	262,390	441,460	256,280	727,231	11,201	29	43
Paraguay	194,400	131,147	287,580	591,779	426,866	820,674	20,880	90	47
French Polynesia	9,879	4,734	20,300	29,763	16,034	55,211	968	13	95
RŽunion	21,329	13,990	31,231	65,378	46,176	89,873		89	91
Rwanda	79,509	35,540	132,872	248,565	124,383	383,373		36	65
Saudi Arabia	152,009	103,604	213,847	468,868	337,790	624,191	1,584	115	61
Sudan	713,990	488,578	1,002,006	2,200,977	1,587,613	2,923,357		116	20
Senegal	314,220	215,380	449,280	962,422	703,216	1,295,220		112	36
Singapore	180,895	38,032	506,530	543,970	153,362	1,338,288	5,631	1	23
Solomon Islands	8,250	4,572	12,834	25,552	15,642	37,037	1	59	107
Sierra Leone	143,041	92,780	212,462	439,787	306,502	607,668		82	59
El Salvador	205,242	141,902	295,506	624,014	458,412	843,641	19,427	110	49
Somalia	114,617	67,014	173,889	354,808	224,893	504,903		71	62
Sao Tome and Principe	5,372	2,326	10,714	16,289	8,203	29,361		14	106
Suriname	19,114	10,933	31,511	57,875	36,560	88,136	201	52	87
Seychelles	2,677	1,375	4,589	8,173	4,772	12,922		40	123
Syrian Arab Republic	91,973	36,485	156,775	289,260	128,876	455,438		26	58
Turks and Caicos Islands	563	230	1,095	1,746	849	3,051	8	15	132
Chad	145,525	85,538	218,800	451,621	286,219	637,502		75	56
Togo	189,659	132,343	268,471	578,418	429,007	770,252		117	54
Thailand	1,903,694	1,373,605	2,621,098	5,823,012	4,424,859	7,596,099	57,589	130	11
Tajikistan	21,693	7,680	39,389	68,289	27,976	113,063		19	83
Tokelau	32	8	76	96	32	205		4	138
Turkmenistan	21,770	8,698	39,089	68,326	31,473	111,794		21	84
Timor-Leste	14,586	8,141	22,488	45,345	27,395	65,380	278	64	97

Tonga	3,494	1,938	5,852	10,584	6,528	16,323	166	46	120
Trinidad and Tobago	48,180	31,084	72,602	145,572	101,375	205,512	1,255	81	79
Tuvalu	155	48	348	488	188	968		6	136
Tanzania (United Republic of)	674,056	456,225	954,435	2,075,066	1,483,724	2,778,664		113	21
Uganda	602,260	395,846	859,315	1,862,790	1,297,641	2,511,403		102	25
Uzbekistan	105,943	45,117	180,283	332,456	158,412	521,236		30	55
Saint Vincent and the Grenadines	3,267	1,994	5,033	9,979	6,659	14,331	63	72	124
Venezuela	866,172	625,770	1,193,623	2,634,742	2,011,455	3,434,002	73,796	129	19
Virgin Islands, British	664	298	1,252	2,014	1,056	3,445	338	20	131
Virgin Islands, U.S.	3,885	1,900	6,977	11,748	6,571	19,306		27	113
Viet Nam	2,603,443	1,890,174	3,578,852	7,965,912	6,081,413	10,371,255	110,217	134	9
Vanuatu	4,274	2,365	6,764	13,222	8,155	19,401	111	58	115
Wallis and Futuna	488	242	860	1,487	843	2,408	4	31	135
Samoa	16,759	9,671	28,084	51,096	32,418	78,612	226	49	90
Yemen	222,930	141,959	324,076	689,860	465,642	945,429	833	86	44
Zambia	148,229	94,049	215,850	458,423	310,077	628,782		83	57
Zimbabwe	80,075	38,569	129,836	250,485	131,542	377,858		45	68

Table T4: Apparent and Inapparent mean and confidence (95%) burden estimates per country. The CI ratio rank is calculated as the ranked index of the apparent confidence interval difference divided by the mean[203]. The CI absolute rank is calculated as the ranked index of the difference in the apparent confidence interval. Only countries with evidence consensus > -25 are included.

E.4 Comparing cartographic and surveillance-based burden estimates

In E.3 we presented two estimates of dengue burden that were different by an order of magnitude. In this section, we reconcile these estimates by discussing and reasonably quantifying each step in the surveillance-based reporting system (Figure SE5).

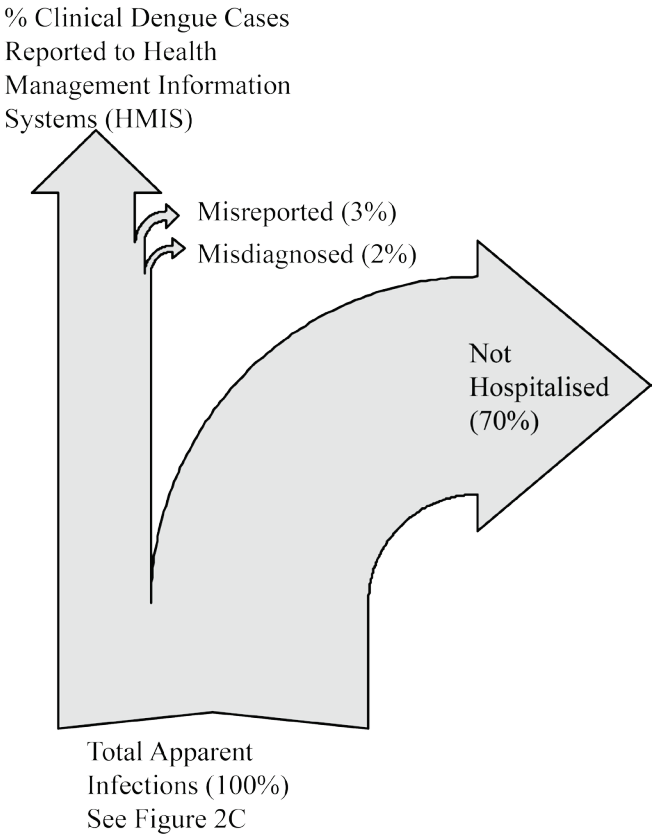

Figure SE5. Hypothetical reporting chain of a dengue virus infection. Of the total number of *apparent infections* only around 30% will seek treatment at official healthcare facilities as opposed to alternative treatment options. Of the formally *hospitalised infections*, a large proportion are misdiagnosed, although total *diagnosed infections* is nearly counter balanced by non-dengue illnesses receiving a dengue diagnosis. Finally, technical, political and logistical barriers exist between the hospital and the governing bodies that result in fewer *reported infections*. These steps collectively ensure that the reported burden of dengue cases is only a small proportion of the total volume of apparent infections. The loss arrows represent an average estimate of the proportion of total apparent infections reported at each level, based on a comparative analysis discussed in E4.

In our estimates we define an apparent dengue infection as any infection that results in visible symptoms, for example nausea or vomiting, rash, aches and pains, mucosal bleeding or restlessness[176], sufficient to disrupt to the individual's daily routine (seeD2.2). Our analysis

suggests a mean figure of 96 million (CI = 67-136) apparent dengue infections per year which captures the complete spectrum of dengue infections from mild to severe. However, of these total apparent infections, there is a wide range in severity[204-207] which affects the treatment-seeking behaviours of infected individuals. The individual may choose to present to formal healthcare facilities operated by Ministries of Health, or the private or non-profit sectors. It is also the case, however, that many care-seekers will choose other options such as homeopathic medicine[208] or ambulatory clinics[157], or will simply use over-the-counter medicines if they seek treatment at all. These decisions depend on a myriad of factors including the availability of the various healthcare facilities and the cost associated with each option. As such, treatment-seeking behaviour for dengue is likely to vary considerably in time (e.g. epidemic versus inter-epidemic periods) and in space (e.g. country to country), as well as by socio-economic, cultural, age and gender groupings. Whilst some of these alternative treatment options may sometimes result in dengue reporting[209], this is likely to be the exception rather than the norm, and in most cases will leave no formal record for inclusion in national health statistics[157 210]. Approximations of the fraction of dengue infections that present to formal healthcare facilities can be made from cohort studies where the hospitalised case numbers are recorded in parallel to the incidence of apparent infections in the general community. From the nine studies in D2.2 that measured both of these parameters[204,205,211-217] an average of 30% (range 18-60) of apparent infections presented to formal healthcare facilities. If generalizable, this would suggest only 28.8 million of the 96 million apparent infections would present to formal healthcare facilities. This is an upper-bound estimate as these studies minimise financial, educational and logistic barriers to healthcare access thus modifying treatment-seeking behaviour. This is an important factor in reconciling these estimates because it accounts for the single biggest loss of dengue infections between total apparent infections and cases reported to the WHO.

Of the apparent dengue infections that do present to an official healthcare facility, a wide spectrum still exists between the mildest and the most severe clinical cases[179,218,219]. This complicates accurate diagnosis of dengue, particularly in the febrile phase where up to 12 other major pathogens may be implicated in the differential diagnosis[179]. This creates two potential problems: (i) clinical diagnosis of dengue infections as other febrile illnesses (under-diagnosis) and (ii) diagnosis of non-dengue febrile illnesses as clinical dengue (over-diagnosis). Under-diagnosis of dengue is common in situations where the disease has a less familiar clinical presentation as compared to other febrile illnesses, for example misdiagnosis as malaria in Africa[220]. Even in cases where dengue is a common diagnosis, often the rigidity of the case definition neglects the broad spectrum and transitory nature of symptoms presented[201,202,221-226]. The extent of under-diagnosis can be estimated by comparing independent attempts to calculate the positive predictive value of a particular clinical dengue case definition. The positive predictive value gives the probability that the case definition used by the hospital will accurately diagnose a true dengue infection. A comparison of case definitions based on the 1997 WHO clinical guidelines[200] for diagnosis of DF revealed an average positive prediction value of 57% (range 16-87)[226-233]. Over-diagnosis of dengue is likely to occur in the later stages of an epidemic when healthcare services are overwhelmed and a diagnosis of dengue is more likely to be suspected over other clinically similar infections[234]. It is also possible to quantify over-diagnosis by observing the proportion of true dengue infections, as determined by laboratory confirmation, out of the total number of clinical dengue diagnoses. This information was extracted from a collection of 29 published clinical outcome studies available on request. We found that, on average, true dengue infections make up only 60% of total clinically diagnosed dengue cases. Of course, over-diagnosis can be reduced when laboratory confirmation is available for clinically-diagnosed

samples; however, this makes up only a small proportion of globally-reported dengue cases (6.9% of total cases reported to WHO regional offices in 2010)[166,196-198]. If we again generalise these figures globally, the suggested 28.8 $\left({}^{30}/_{100} \times 96\right)$ million presenting to official healthcare facilities is further reduced to 16.4 $\left({}^{57}/_{100} \times 28.8\right)$ million in light of under-diagnosis, 27.4$\left({}^{100}/_{60} \times 16.4\right)$ if over-diagnosis is included as well, and 26.6 $((1 - 0.069) \times 27.4 + 0.6 \times 0.069 \times 27.4)$ million taking into account both under and over-diagnosis with laboratory confirmation available for 6.9% of clinical samples.

Once a clinical or laboratory diagnosis of dengue has been given, the remaining step in the reporting pathway is for this case to be incorporated in the national health management information systems (HMIS) and then to the WHO. The systems in place to transfer these data from the hospital to the WHO vary widely in their use of technology and frequency of reporting. While many health system records are now computer-based, much of the primary reporting is still gathered over the telephone or by fax[235] which creates inherent standardisation issues. Furthermore, gaps in weekly reporting schedules can be observed during national holidays or during temporary shortages of available staff and/or funding. These logistical errors are compounded by political conflicts of interest in what is reported and how it is used to avoid accusations of liability. Thus, communication through HMIS systems will inevitably result in minor data losses. Although a lack of available data makes this step difficult to quantify, we were able to estimate the data loss at the final step of HMIS reporting to the WHO. A comparison of reported DF cases on ministry of health (MoH) websites with those reported to the WHO from 2009-2011 found that 95% of total cases were reported to the WHO [166,196-198]. No persistently under-reported countries were identified, but one-off years where data was not reported were observed. If we assume a 95% reporting capacity across *both* steps (local hospital to MoH and MoH to WHO) our suggested figure of

26.6 million diagnosed dengue infections would translate to 24.0 $\left(26.5 \times {}^{95}/_{100} \times {}^{95}/_{100}\right)$ million potentially reportable cases per year or, overall approximately one quarter of the total apparent dengue infections that occur each year.

This figure of 24.0 million potentially reported cases is an absolute upper bound of what we might expect to be reported to the WHO globally. This is because the studies we have used all estimate the upper bound of reportable cases. The hospitalisation rates are heavily influenced by increased healthcare provision. Case-definition positive predictive values are calculated under controlled situations that do not accurately reflect the strains on the healthcare system that vary both through space (e.g. equipment deficits in rural areas) and time (e.g. resource constraints during epidemics). Furthermore, estimates of over-diagnoses often take a very broad, and hence less specific, clinical definition of dengue so as to maximise study participant number; this increases the number of falsely clinically-diagnosed dengue patients. Finally, the losses through HMIS and between HMIS and the WHO were calculated from countries that have already been shown to proactively use their reported data (in so far as they display it on their MoH website). In countries where there is no motivation to use this data at a national level, it is easy to see how further data losses could occur, meaning an adjustment factor of only 5% loss at this stage is likely to be extremely conservative.

PAHO, with 76% of total reported dengue cases in 2010, provides the most accurate and consistently reported dengue data (Table S1). PAHO reported 1.7 million dengue cases in 2010 [166]. For this region our cartographic burden-estimation approach predicts 13.3 million apparent dengue infections suggesting a clinical burden of 3.99 million cases, 3.68 million clinical dengue diagnoses (with over and under reporting and laboratory confirmation) and

3.32 million cases reported to the WHO. The potential for over-estimation of parameters discussed above, in particular the over-diagnosis of dengue due to other febrile illnesses, is likely to be a specific source of over-estimation in the Americas. Considering this, parameter overestimation of as little as 10% is enough to reduce our burden estimate to levels comparable with WHO regional office reports (1.77 million compared to 1.7 million).

In summary, the aim of this section is not to quantitatively estimate the number of cases at each stage in the healthcare system, but rather to show that our cartographic burden estimates of apparent infections are plausible when compared to surveillance-based estimates of clinical dengue cases once the inevitable underreporting of the latter approach is considered. We have shown that the biggest under-reporting step between apparent infections and reported cases occurs during treatment-seeking. Further losses occur due to difficulties in diagnosis and through errors in reporting. A detailed understanding of the link between total apparent infections and total reported cases is an important consideration if the true burden of dengue is to be estimated at various levels.

F: References

1 Freifeld, C. C., Mandl, K. D., Reis, B. Y. & Brownstein, J. S. HealthMap: global infectious disease monitoring through automated classification and visualization of internet media reports. *J. Am. Med. Inform. Assoc.* **15**, 150-157, (2008).

2 Brady, O. J. *et al.* Refining the global spatial limits of dengue transmission in 2012 by evidence-based consensus. *PLoS Negl. Trop. Dis.* **6**, e1760, (2012).

3 Simmons, C. P., Farrar, J. J., van Vinh Chau, N. & Wills, B. Dengue. *N. Engl. J. Med.* **366**, 1423-1432, (2012).

4 Cardosa, J. *et al.* Dengue virus serotype 2 from a sylvatic lineage isolated from a patient with dengue hemorrhagic fever. *PLoS Negl. Trop. Dis.* **3**, e423, (2009).

5 Gubler, D. J. & Kuno, G. *Dengue and dengue hemorrhagic fever*. (Cab International, 1997).

6 Gubler, D. J. Dengue and dengue hemorrhagic fever. *Clin. Microbiol. Rev.* **11**, 480-496, (1998).

7 Hales, S., De Wet, N., Maindonald, J. & Woodward, A. Potential effect of population and climate changes on global distribution of dengue fever: an empirical model. *Lancet* **360**, 830-834, (2002).

8 Patz, J. A., Martens, W., Focks, D. A. & Jetten, T. H. Dengue fever epidemic potential as projected by general circulation models of global climate change. *Environ. Health Perspect.* **106**, 147, (1998).

9 Barbazan, P. *et al.* Modelling the effect of temperature on transmission of dengue. *Med. Vet. Entomol.* **24**, 66-73, (2010).

10 Ooi, E. E. & Gubler, D. J. Global spread of epidemic dengue: the influence of environmental change. *Future. Virol.* **4**, 571-580, (2009).

11 Degallier, N. *et al.* Toward an early warning system for dengue prevention: modeling climate impact on dengue transmission. *Clim. Change* **98**, 581-592, (2010).

12 Martens, W. J. M., Jetten, T. H. & Focks, D. A. Sensitivity of malaria, schistosomiasis and dengue to global warming. *Clim. Change* **35**, 145-156, (1997).

13 Halstead, S. B. Dengue virus-mosquito interactions. *Annu. Rev. Entomol.* **53**, 273-291, (2008).

14 Banu, S., Hu, W., Hurst, C. & Tong, S. Dengue transmission in the Asia-Pacific region: impact of climate change and socio-environmental factors. *Trop. Med. Int. Health* **16**, 598-607, (2011).

15 Jansen, C. C. & Beebe, N. W. The dengue vector *Aedes aegypti*: what comes next. *Microbes Infect.* **12**, 272-279, (2010).

16 Wilder-Smith, A. & Gubler, D. J. Geographic expansion of dengue: the impact of international travel. *Med. Clin. N. Am.* **92**, 1377-1390, (2008).

17 Romero-Vivas, C. M. & Falconar, A. K. Investigation of relationships between *Aedes aegypti* egg, larvae, pupae, and adult density indices where their main breeding sites were located indoors. *J. Am. Mosq. Control Assoc.* **21**, 15-21, (2005).

18 Johansson, M. A., Dominici, F. & Glass, G. E. Local and global effects of climate on dengue transmission in Puerto Rico. *PLoS Negl. Trop. Dis.* **3**, e382, (2009).

19 Li, C. F., Lim, T. W., Han, L. L. & Fang, R. Rainfall, abundance of *Aedes aegypti* and dengue infection in Selangor, Malaysia. *Southeast Asian J. Trop. Med. Public Health* **16**, 560-560, (1985).

20 Hurtado-Diaz, M., Riojas-Rodriguez, H., Rothenberg, S. J., Gomez-Dantes, H. & Cifuentes, E. Short communication: impact of climate variability on the incidence of dengue in Mexico. *Trop. Med. Int. Health* **12**, 1327-1337, (2007).

21 Eamchan, P., Nisalak, A., Foy, H. M. & Chareonsook, O. A. Epidemiology and control of dengue virus infections in Thai villages in 1987. *Am. J. Trop. Med. Hyg.* **41**, 95-101, (1989).

22 Aiken, S. R., Frost, D. B. & Leigh, C. H. Dengue hemorrhagic fever rainfall in Penninsular Malyasia: some suggested relationships. *Soc. Sci. Med.* **14D**, 307-316, (1980).

23 Wiwanitkit, V. An observation on correlation between rainfall and the prevalence of clinical cases of dengue in Thailand. *J. Vector Borne Dis.* **43**, 73, (2006).

24 Heng, B., Goh, K. & Neo, K. *Environmental temperature, Aedes aegypti house index and rainfall as predictors of annual epidemics of dengue fever and dengue haemorrhagic fever in Singapore*. Vol. Dengue in Singapore (CAB international, 1998).

25 Tun-Lin, W., Burkot, T. R. & Kay, B. H. Effects of temperature and larval diet on development rates and survival of the dengue vector *Aedes aegypti* in north Queensland, Australia. *Med. Vet. Entomol.* **14**, 31-37, (2000).

26 Delatte, H., Gimonneau, G., Triboire, A. & Fontenille, D. Influence of temperature on immature development, survival, longevity, fecundity, and gonotrophic cycles of *Aedes albopictus*, vector of chikungunya and dengue in the Indian Ocean. *J. Med. Entomol.* **46**, 33-41, (2009).

27 McLean, D. M. *et al.* Vector capability of *Aedes aegypti* mosquitoes for California encephalitis and dengue viruses at various temperatures. *Can. J. Microbiol.* **20**, 255-262, (1974).

28 Watts, D. M., Burke, D. S., Harrison, B. A., Whitmire, R. E. & Nisalak, A. Effect of temperature on the vector efficiency of *Aedes aegypti* for dengue 2 virus. (DTIC Document, 1986).

29 Chowell, G., Cazelles, B., Broutin, H. & Munayco, C. V. The influence of geographic and climate factors on the timing of dengue epidemics in Peru, 1994-2008. *BMC Infect. Dis.* **11**, 164, (2011).

30 Pinto, E., Coelho, M., Oliver, L. & Massad, E. The influence of climate variables on dengue in Singapore. *Int. J. Environ. Health Res.* **21**, 415-426, (2011).

31 Raheel, U. *et al.* Dengue fever in the Indian subcontinent: an overview. *J. Infect. Dev. Ctries.* **5**, 239-247, (2011).

32 Focks, D., Haile, D., Daniels, E. & Mount, G. Dynamic life table model for *Aedes aegypti* (Diptera: Culicidae): analysis of the literature and model development. *J. Med. Entomol.* **30**, 1003-1017, (1993).

33 Focks, D., Haile, D., Daniels, E. & Mount, G. Dynamic life table model for *Aedes aegypti* (Diptera: Culicidae): simulation results and validation. *J. Med. Entomol.* **30**, 1018-1028, (1993).

34 Gething, P. W. *et al.* Modelling the global constraints of temperature on transmission of *Plasmodium falciparum* and *P. vivax*. *Parasit. Vectors* **4**, 92-92, (2011).

35 Linthicum, K. J. *et al.* Climate and satellite indicators to forecast Rift Valley fever epidemics in Kenya. *Science* **285**, 397-400, (1999).

36 Cox, J., Grillet, M. E., Ramos, O. M., Amador, M. & Barrera, R. Habitat segregation of dengue vectors along an urban environmental gradient. *Am. J. Trop. Med. Hyg.* **76**, 820-826, (2007).

37 Sota, T. & Mogi, M. Interspecific variation in desiccation survival time of *Aedes (Stegomyia)* mosquito eggs is correlated with habitat and egg size. *Oecologia* **90**, 353-358, (1992).

38 Reiskind, M. & Lounibos, L. Effects of intraspecific larval competition on adult longevity in the mosquitoes *Aedes aegypti* and *Aedes albopictus*. *Med. Vet. Entomol.* **23**, 62-68, (2009).

39 Costa, E. A. P. A., Santos, E. M. M., Correia, J. C. & Albuquerque, C. M. R. Impact of small variations in temperature and humidity on the reproductive activity and survival of *Aedes aegypti* (*Diptera, Culicidae)*. *Rev. Bras. Entomol.* **54**, 488-493, (2010).

40 Luz, C., Tai, M., Santos, A. & Silva, H. Impact of moisture on survival of *Aedes aegypti* eggs and ovicidal activity of Metarhizium anisopliae under laboratory conditions. *Mem. Inst. Oswaldo Cruz* **103**, 214-215, (2008).

41 Russell, B., Kay, B. & Shipton, W. Survival of *Aedes aegypti* (Diptera: Culicidae) eggs in surface and subterranean breeding sites during the Northern Queensland dry season. *J. Med. Entomol.* **38**, 441-445, (2001).

42 Trpis, M. Dry season survival of *Aedes aegypti* eggs in various breeding sites in the Dar es Salaam area, Tanzania. *Bull World Health Organ* **47**, 433, (1972).

43 Fuller, D., Troyo, A. & Beier, J. El Nino southern oscillation and vegetation dynamics as predictors of dengue fever cases in Costa Rica. *Env. Res. Lett.* **4**, 014011, (2009).

44 Troyo, A., Fuller, D. O., Calderon- Arguedas, O., Solano, M. E. & Beier, J. C. Urban structure and dengue incidence in Puntarenas, Costa Rica. *Singapore J. Trop. Med.* **30**, 265-282, (2009).

45 Bisset Lazcano, J. A. *et al.* Ecological factors linked to the presence of *Aedes aegypti* larvae in highly infested areas of Playa, a municipality belonging to Ciudad de La Habana, Cuba. *Rev. Panam. Salud Publica* **19**, 379-384, (2006).

46 Barrera, R., Amador, M. & Clark, G. G. Use of the pupal survey technique for measuring *Aedes aegypti* (*Diptera: Culicidae)* productivity in Puerto Rico. *Am. J. Trop. Med. Hyg.* **74**, 290-302, (2006).

47 Mena, N., Troyo, A., Bonilla-Carrion, R. & Calderon-Arguedas, O. Factors associated with incidence of dengue in Costa Rica. *Rev. Panam. Salud. Publica.* **29**, 234-242, (2011).

48 Flauzino, R. F. *et al.* Spatial heterogeneity of dengue fever in local studies, City of Niteroi, Southeastern Brazil. *Rev. Saude Publica* **43**, 1035-1043, (2009).

49 Ratho, R. K., Mishra, B., Kaur, J., Kakkar, N. & Sharma, K. An outbreak of dengue fever in periurban slums of Chandigarh, India, with special reference to entomological and climatic factors. *Indian J. Med. Sci.* **59**, 518-526, (2005).

50 Lifson, A. R. Mosquitoes, models, and dengue. *Lancet* **347**, 1201-1202, (1996).

51 Schmidt, W. P. *et al.* Population density, water supply, and the risk of dengue fever in Vietnam: cohort study and spatial analysis. *PLoS Med.* **8**, e1001082, (2011).

52 Liebman, K. A. *et al.* Spatial dimensions of dengue virus transmission across interepidemic and epidemic periods in Iquitos, Peru (1999-2003). *PLoS Negl. Trop. Dis.* **6**, e1472, (2012).

53 Gallup, J. L., Sachs, J. D. & Mellinger, A. D. Geography and economic development. *NEBR Working Paper Series* **No. 6849**, (1998).

54 Rinaldi, P. N. Epidemiologic risk of dengue and the role of human movement in an economically disadvantaged urban environment. *Emory Electronic Thesis*, (2011).

55 Adams, B. & Kapan, D. D. Man Bites Mosquito: Understanding the Contribution of Human Movement to Vector-Borne Disease Dynamics. *PLoS One* **4**, e6763-e6763, (2009).

56 Gubler, D. J. Dengue and dengue hemorrhagic fever: its history and resurgence as a global public health problem. *Dengue and dengue hemorrhagic fever*, 1-22, (1997).

57 Hollingsworth, T. D., Ferguson, N. M. & Anderson, R. M. Frequent travelers and rate of spread of epidemics. *Emerg. Infect. Dis.* **13**, 1288, (2007).

58 Tatem, A. J., Hay, S. I. & Rogers, D. J. Global traffic and disease vector dispersal. *Proc. Natl. Acad. Sci.* **103**, 6242-6247, (2006).

59 Harrington, L. C. *et al.* Dispersal of the dengue vector *Aedes aegypti* within and between rural communities. *Am. J. Trop. Med. Hyg.* **72**, 209-220, (2005).

60 Nelson, A. Estimated travel time to the nearest city of 50,000 or more people in year 2000. *Global Environment Monitoring Unit, Joint Research Centre of the European Commission*, (2008).

61 Joint Research Center Global Environmental Modelling Unit. Travel time to major cities: A global map of Accessibility. http://bioval.jrc.ec.europa.eu/products/gam/sources.htm.

62 Chakravarti, A., Arora, R. & Luxemburger, C. Fifty years of dengue in India. *Trans. R. Soc. Trop. Med. Hyg.*, (2012).

63 Cummings, D. A. *et al.* Travelling waves in the occurrence of dengue haemorrhagic fever in Thailand. *Nature* **427**, 344-347, (2004).

64 Almeida, M. C. D., Caiaffa, W. T., Assuncao, R. M. & Proietti, F. A. Spatial vulnerability to dengue in a Brazilian urban area during a 7-year surveillance. *J. Urban Health* **84**, 334-345, (2007).

65 Ahmed, S. *et al.* Dengue fever outbreak in Karachi 2006--a study of profile and outcome of children under 15 years of age. *JPMA* **58**, 4, (2008).

66 Nagi, A. G., Murad, R. & Baig, M. Dengue fever outbreak among children in Karachi: experience at a tertiary care children hospital. *JBUMDC ISSN 2220-7562*, 44, (2011).

67 Heddini, A., Janzon, R. & Linde, A. Increased number of dengue cases in Swedish travellers to Thailand. *J. Infect. Dis.* **195**, 1089-1096, (2007).

68 Balk, D. L. *et al.* Determining global population distribution: methods, applications and data. *Adv. Parasitol.* **62**, 119-156, (2006).

69 Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965-1978, (2005).

70 Hutchinson, M. F. Interpolating mean rainfall using thin-plate smoothing splines. *Int. J. Geogr. Inf. Sci.* **9**, 385-403, (1995).

71 Scharlemann, J. P. W. *et al.* Global data for ecology and epidemiology: a novel algorithm for temporal Fourier processing MODIS data. *PLoS One* **3**, e1408, (2008).

72 Rogers, D. J., Hay, S. I. & Packer, M. J. Predicting the distribution of tsetse flies in West Africa using temporal Fourier processed meteorological satellite data. *Ann. Trop. Med. Parasitol.* **90**, 225-241, (1996).

73 Paaijmans, K. P. *et al.* Influence of climate on malaria transmission depends on daily temperature variation. *Proc. Natl. Acad. Sci.* **107**, 15135-15139, (2010).

74 Paaijmans, K. P., Read, A. F. & Thomas, M. B. Understanding the link between malaria risk and climate. *Proc. Natl. Acad. Sci.* **106**, 13844-13849, (2009).

75 Hay, S. I. An overview of remote sensing and geodesy for epidemiology and public health application. *Adv. Parasitol.* **47**, 1-35, (2000).

76 Hay, S. I., Tatem, A. J., Graham, A. J., Goetz, S. J. & Rogers, D. J. Global environmental data for mapping infectious disease distribution. *Adv. Parasitol.* **62**, 37-77, (2006).

77 Elvidge, C. D. *et al.* Radiance calibration of DMSP-OLS low-light imaging data of human settlements. *Remote Sens. Environ.* **68**, 77-88, (1999).
78 Elvidge, C. D. *et al.* in *Remotely Sensed Cities* (ed V. Mesev) 281-333 (Taylor and Francis, 2003).
79 Tatem, A. J., Noor, A. M. & Hay, S. I. Assessing the accuracy of satellite derived global and national urban maps in Kenya. *Remote Sens. Environ.* **96**, 87-97, (2005).
80 Gething, P. W. *et al.* A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malar. J.* **10**, 378, (2011).
81 Hay, S. I. *et al.* A world malaria map: *Plasmodium falciparum* endemicity in 2007. *PLoS Med.* **6**, e1000048, (2009).
82 Hay, S. I., Guerra, C. A., Tatem, A. J., Atkinson, P. M. & Snow, R. W. Urbanization, malaria transmission and disease burden in Africa. *Nat. Rev. Microbiol.* **3**, 81-90, (2005).
83 Hay, S. I., Noor, A. M., Nelson, A. & Tatem, A. J. The accuracy of human population maps for public health application. *Trop. Med. Int. Health* **10**, 1073-1086, (2005).
84 Balk, D., Pozzi, F., Yetman, G., Deichmann, U. & Nelson, A. The distribution of people and the dimension of place: methodologies to improve the global estimation of urban extents. *Draft version. Palisades, Columbia University: New York, NY, CIESIN*, (2004).
85 Nordhaus, W. New metrics for environmental economics: gridded economic data. *Integrated Assessment* **8**, (2008).
86 Nordhaus, W. D. Geography and macroeconomics: new data and new findings. *Proc. Natl. Acad. Sci.* **103**, 3510-3517, (2006).
87 Dormann, C. F. *et al.* Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, (2012).
88 Prendergast, J. R., Quinn, R. M. & Lawton, J. H. The gaps between theory and practice in selecting nature reserves. *Conserv. Biol.* **13**, 484-492, (1999).
89 Stevens, K. B. & Pfeiffer, D. U. Spatial modelling of disease using data-and knowledge-driven approaches. *Spat. Spattemporal. Epidemiol.*, (2011).
90 Guisan, A. & Zimmermann, N. E. Predictive habitat distribution models in ecology. *Ecol. Model.* **135**, 147-186, (2000).
91 Zaniewski, A. E., Lehmann, A. & Overton, J. M. C. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecol. Model.* **157**, 261-280, (2002).
92 Busby, J. BIOCLIM-a bioclimate analysis and prediction system. *Plant Prot. Q.* **6**, (1991).
93 Austin, M. P. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecol. Model.* **157**, 101-118, (2002).
94 Elith, J. *et al.* Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29**, 129-151, (2006).
95 Phillips, S. J. *et al.* Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecol. Appl.* **19**, 181-197, (2009).
96 Hernandez, P. A., Graham, C. H., Master, L. L. & Albert, D. L. The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography* **29**, 773-785, (2006).
97 Mateo, R. G., Croat, T. B., Felicisimo, A. M. & Munoz, J. Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections. *Divers. Distrib.* **16**, 84-94, (2010).

98 McCullagh, P. & Nelder, J. A. *Generalized linear models*. (Chapman & Hall/CRC, 1989).
99 Yee, T. W. & Mitchell, N. D. Generalized additive models in plant ecology. *J. Veg. Sci.* **2**, 587-602, (1991).
100 Hastie, T., Tibshirani, R. & Friedman, J. H. *The elements of statistical learning*. (Springer, 2009).
101 Friedman, J. H. & Meulman, J. J. Multiple additive regression trees with application in epidemiology. *Stat. Med.* **22**, 1365-1381, (2003).
102 Breiman, L. *Classification and regression trees*. (Chapman & Hall/CRC, 1984).
103 Phillips, S. J. & Dudík, M. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* **31**, 161-175, (2008).
104 Jaynes, E. T. Information theory and statistical mechanics. II. *Phys. Rev.* **108**, 171, (1957).
105 Skilling, J. Data analysis: the maximum entropy method. *Nature* **309**, 748-749, (1984).
106 Elith, J., Leathwick, J. R. & Hastie, T. A working guide to boosted regression trees. *J. Anim. Ecol.* **77**, 802-813, (2008).
107 Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.*, 1189-1232, (2001).
108 De'Ath, G. Boosted trees for ecological modeling and prediction. *Ecology* **88**, 243-251, (2007).
109 Moffett, A., Shackelford, N. & Sarkar, S. Malaria in Africa: vector species' niche models and relative risk maps. *PLoS One* **2**, e824, (2007).
110 VanDerWal, J., Shoo, L. P., Graham, C. & Williams, S. E. Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecol. Model.* **220**, 589-594, (2009).
111 Chefaoui, R. M. & Lobo, J. M. Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecol. Model.* **210**, 478-486, (2008).
112 Bishop, C. M. & ligne, S. *Pattern recognition and machine learning*. Vol. 4 (springer New York, 2006).
113 Hastie, T., Tibshirani, R. & Friedman, J. *The elements of statistical learning: data mining, inference and prediction*. Second edn, (Springer, 2009).
114 Ridgeway, G. gbm: Generalized boosted regression models. R package, version 1.3-5. *RAND Statistics Group, Santa Monica, California*, (2006).
115 Rogers, D. Models for vectors and vector-borne diseases. *Adv. Parasitol.* **62**, 1-35, (2006).
116 Allouche, O., Tsoar, A. & Kadmon, R. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *J. Appl. Ecol.* **43**, 1223-1232, (2006).
117 Youden, W. Index for rating diagnostic tests. *Cancer* **3**, 32-35, (1950).
118 Fleiss, J. L. & Cohen, J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. Psychol. Meas.*, (1973).
119 Fleiss, J. L., Levin, B. & Paik, M. C. *The measurement of interrater agreement*. Third edn, (John Wiley & Sons, 2004).
120 Freeman, E. A. & Moisen, G. PresenceAbsence: An R Package for Presence Absence Analysis. *J. Stat. Soft.* **23**, 1-31, (2008).
121 Hijmans, R. J. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. *Ecology* **93**, 679-688, (2012).
122 Hijmans, R. J., Phillips, S., Leathwick, J. & Elith, J. Package DISMO. *Circles* **9**, 1, (2011).

123 Engler, R., Guisan, A. & Rechsteiner, L. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *J. Appl. Ecol.* **41**, 263-274, (2004).
124 Barbet-Massin, M., Jiguet, F., Albert, C. H. & Thuiller, W. Selecting pseudo-absences for species distribution models: how, where and how many? *Methods Ecol. Evol.*, (2011).
125 Stokland, J. N. & Halvorsen, R. Species distribution modelling--Effect of design and sample size of pseudo-absence observations. *Ecol. Model.*, (2011).
126 Wisz, M. & Guisan, A. Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. *BMC Ecol.* **9**, 8, (2009).
127 Warton, D. I. & Shepherd, L. C. Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *Ann. Appl. Stat.* **4**, 1383-1402, (2010).
128 Thuiller, W. Patterns and uncertainties of species' range shifts under climate change. *Glob. Chang. Biol.* **10**, 2020-2027, (2004).
129 Ward, G., Hastie, T., Barry, S., Elith, J. & Leathwick, J. R. Presence-Only Data and the EM Algorithm. *Biometrics* **65**, 554-563, (2009).
130 Phillips, S. J. & Elith, J. Logistic methods for resource selection functions and presence-only species distribution models. *AAAI (Association for the Advancement of Artificial Intelligence), San Francisco, USA*, (2011).
131 Lobo, J. M. & Tognelli, M. F. Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. *J. Nat. Conserv.* **19**, 1-7, (2011).
132 Thuiller, W., Brotons, L., Araújo, M. B. & Lavorel, S. Effects of restricting environmental range of data to project current and future species distributions. *Ecography* **27**, 165-172, (2004).
133 Lobo, J., Verdú, J. & Numa, C. Environmental and geographical factors affecting the Iberian distribution of flightless Jekelius species (Coleoptera: Geotrupidae). *Divers. Distrib.* **12**, 179-188, (2006).
134 Hirzel, A. H., Helfer, V. & Metral, F. Assessing habitat-suitability models with a virtual species. *Ecol. Model.* **145**, 111-121, (2001).
135 Rogers, D. J., Wilson, A. J., Hay, S. I. & Graham, A. J. The global distribution of yellow fever and dengue. *Adv. Parasitol.* **62**, 181-220, (2006).
136 Sinka, M. E. *et al.* The dominant Anopheles vectors of human malaria in the Americas: occurrence data, distribution maps and bionomic precis. *Parasit. Vectors* **3**, 72-72, (2010).
137 Keating, K. A. & Cherry, S. Use and interpretation of logistic regression in habitat-selection studies. *J. Wildl. Manag.* **68**, 774-789, (2004).
138 Lancaster, T. & Imbens, G. Case-control studies with contaminated controls. *J. Econom.* **71**, 145-160, (1996).
139 McLachlan, G. J. & Krishnan, T. *The EM algorithm and extensions*. Vol. 274 (Wiley New York, 1997).
140 Pearce, J. L. & Boyce, M. S. Modelling distribution and abundance with presence-only data. *J. Appl. Ecol.* **43**, 405-412, (2006).
141 Araújo, M. B. & New, M. Ensemble forecasting of species distributions. *Trends Ecol. Evol.* **22**, 42-47, (2007).
142 Buisson, L., Thuiller, W., Casajus, N., Lek, S. & Grenouillet, G. Uncertainty in ensemble forecasting of species distribution. *Glob. Chang. Biol.* **16**, 1145-1157, (2010).

143 Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R. K. & Thuiller, W. Evaluation of consensus methods in predictive species distribution modelling. *Divers. Distrib.* **15**, 59-69, (2009).

144 Bates, J. M. & Granger, C. W. J. The combination of forecasts. *OR*, 451-468, (1969).

145 WHO. *Dengue and dengue Haemorrhagic fever. Fact sheet no.117.*, <http://www.who.int/mediacentre/factsheets/fs117/en/> (2012).

146 Guzman, M. G. & Kouri, G. Dengue: an update. *Lancet Infect. Dis.* **2**, 33-42, (2002).

147 Gibbons, R. V. & Vaughn, D. W. Dengue: an escalating problem. *BMJ* **324**, 1563-1566, (2002).

148 Halstead, S. B. Pathogenesis of dengue: challenges to molecular biology. *Science* **239**, 476-481, (1988).

149 Monath, T. P. Yellow fever and dengue-the interactions of virus, vector and host in the re-emergence of epidemic disease. *Semin. Virol.* **5**, 133-145, (1994).

150 Rigau-Perez, J. G. *et al.* Dengue and dengue haemorrhagic fever. *Lancet* **352**, 971-977, (1998).

151 Rodhain, F. La situation de la dengue dans le monde. *Bull. Soc. Pathol. Exot.* **89**, 87-90, (1996).

152 Suaya, J., Shepard, D., Beatty, M. Dengue: Burden Of Disease And Costs Of Illness. *Scientific Working Group on Dengue Research (Vol. TDR/SWG/08)* (2007).

153 LeDuc, J., Esteves, K., Gratz, N. in *The Global Epidemiology of Infectious Diseases* (ed C. Murray, Lopez, A., Mathers, C) (World Health Organization, 2004).

154 Cattand, P. *et al.* in *Disease Control Priorities in Developing Countries* (eds D. T. Jamison *et al.*) (2006).

155 WHO. *The Global Burden of Disease: 2004 update.*, <http://www.who.int/healthinfo/global_burden_disease/GBD_report_2004update_full.pdf> (2004).

156 Carrasco, L. R. *et al.* Economic impact of dengue illness and the cost-effectiveness of future vaccination programs in Singapore. *PLoS Negl. Trop. Dis.* **5**, e1426, (2011).

157 Wichmann, O. *et al.* Dengue in Thailand and Cambodia: an assessment of the degree of underrecognized disease burden based on reported cases. *PLoS Negl. Trop. Dis.* **5**, e996, (2011).

158 Shepard, D. S., Coudeville, L., Halasa, Y. A., Zambrano, B. & Dayan, G. H. Economic impact of dengue illness in the Americas. *Am. J. Trop. Med. Hyg.* **84**, 200-207, (2011).

159 Standish, K., Kuan, G., Aviles, W., Balmaseda, A. & Harris, E. High dengue case capture rate in four years of a cohort study in Nicaragua compared to national surveillance data. *PLoS Negl. Trop. Dis.* **4**, e633, (2010).

160 Suaya, J. A. *et al.* Cost of dengue cases in eight countries in the Americas and Asia: a prospective study. *Am. J. Trop. Med. Hyg.* **80**, 846-855, (2009).

161 Garg, P., Nagpal, J., Khairnar, P. & Seneviratne, S. L. Economic burden of dengue infections in India. *Trans. R. Soc. Trop. Med. Hyg.* **102**, 570-577, (2008).

162 Clark, D. V., Mammen, M. P., Jr., Nisalak, A., Puthimethee, V. & Endy, T. P. Economic impact of dengue fever/dengue hemorrhagic fever in Thailand at the family and population levels. *Am. J. Trop. Med. Hyg.* **72**, 786-791, (2005).

163 Luz, P. M., Grinsztejn, B. & Galvani, A. P. Disability adjusted life years lost to dengue in Brazil. *Trop. Med. Int. Health* **14**, 237-246, (2009).

164 Meltzer, M. I., Rigau-Perez, J. G., Clark, G. G., Reiter, P. & Gubler, D. J. Using disability-adjusted life years to assess the economic impact of dengue in Puerto Rico: 1984-1994. *Am. J. Trop. Med. Hyg.* **59**, 265-271, (1998).

165 Gething, P. W. *et al.* Estimating the number of paediatric fevers associated with malaria infection presenting to Africa's public health sector in 2007. *PLoS Med.* **7**, e1000301, (2010).

166 WHO. *Pan American Health Organization (PAHO) website*, <http://new.paho.org/hq/index.php?option=com_content&task=view&id=264&Itemid=363&lang=es> (2012).

167 Armien, B. *et al.* Clinical characteristics and national economic cost of the 2005 dengue epidemic in Panama. *Am. J. Trop. Med. Hyg.* **79**, 364-371, (2008).

168 Beatty, M., Letson, G.W., Margolis, H.S. Estimating the global burden of dengue. *Am. J. Trop. Med. Hyg.* **81**, 231, (2009).

169 Beatty, M. E., Letson, V.W., Margolis, H.S. in *2nd International Conference on Dengue and Dengue Haemorrhagic Fever.* (Phuket, Thailand, 2009).

170 Patil, A. P. *et al.* Defining the relationship between *Plasmodium falciparum* parasite rate and clinical disease: statistical models for disease burden estimation. *Malar. J.* **8**, 186-186, (2009).

171 Snow, R. W., Guerra, C. A., Noor, A. M., Myint, H. Y. & Hay, S. I. The global distribution of clinical episodes of *Plasmodium falciparum* malaria. *Nature* **434**, 214-217, (2005).

172 Monath, T. P. Yellow fever and dengue-the interactions of virus, vector and host in the re-emergence of epidemic disease. *Sem Virol* **5**, 133-145, (1994).

173 Rodhain, F. La situation de la dengue dans le monde. *Bull Soc Pathol Exot* **89**, 87-90, (1996).

174 TDR/W.H.O. *Report of the Scientific Working Group on Dengue, 2006. TDR/SWG/08*. (TDR/World Health Organization, 2006).

175 Beatty, M. E., Letson, G. W. & Margolis, H. S. Estimating the global burden of dengue. *Am. J. Trop. Med. Hyg.* **81**, 231, (2009).

176 WHO. *Dengue guidelines for diagnosis, treatment, prevention and control*, <http://whqlibdoc.who.int/publications/2009/9789241547871_eng.pdf> (2009).

177 Endy, T. P. *et al.* Determinants of inapparent and symptomatic dengue infection in a prospective study of primary school children in Kamphaeng Phet, Thailand. *PLoS Negl. Trop. Dis.* **5**, e975, (2011).

178 Kao, C. L., King, C. C., Chao, D. Y., Wu, H. L. & Chang, G. J. Laboratory diagnosis of dengue virus infection: current and future perspectives in clinical diagnosis and public health. *J. Microbiol. Immunol. Infect.* **38**, 5-16, (2005).

179 Simmons, C., Farrar, J., Chau, N., Wills, B. Dengue. *N. Engl. J. Med.* **366**, (2012).

180 Cuzzubbo, A. J. *et al.* Comparison of PanBio Dengue Duo IgM and IgG capture ELISA and venture technologies dengue IgM and IgG dot blot. *J. Clin. Virol.* **16**, 135-144, (2000).

181 Innis, B. L. in *Dengue and Dengue Haemorrhagic Fever* (ed D. Gubler, Kuno, G.) 221-243 (CAB International, 1997).

182 Allwinn, R., Doerr, H. W., Emmerich, P., Schmitz, H. & Preiser, W. Cross-reactivity in flavivirus serology: new implications of an old finding? *Med. Microbiol. Immunol.* **190**, 199-202, (2002).

183 Vong, S. *et al.* Under-recognition and reporting of dengue in Cambodia: a capture-recapture analysis of the National Dengue Surveillance System. *Epidemiol. Infect.* **140**, 491-499, (2012).

184 Chungue, E., Boutin, J. P. & Roux, J. Intérêt du Titrage des IgM par Technique Immunoenzymatique Pour le Sérodiagnostic et la Surveillance Épidémiologique de la Dengue en Polynésie Française. *Res. Virol.* **140**, 229-240, (1989).

185 Teixeira Mda, G. *et al.* Dynamics of dengue virus circulation: a silent epidemic in a complex urban area. *Trop. Med. Int. Health* **7**, 757-762, (2002).
186 Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian data analysis*. (CRC press, 2004).
187 Cameron, A. C. & Trivedi, P. K. Econometric models based on count data. Comparisons and applications of some estimators and tests. *Journal of applied econometrics* **1**, 29-53, (1986).
188 Winkelmann, R. *Econometric analysis of count data*. (Springer Verlag, 2008).
189 Hilbe, J. M. *Negative binomial regression*. (Cambridge Univ Pr, 2011).
190 Rasmussen, C. Gaussian processes in machine learning. *Lect. Notes. Artiff. Int.*, 63-71, (2004).
191 Banerjee, S., Carlin, B. P. & Gelfand, A. E. *Hierarchical modeling and analysis for spatial data*. Vol. 101 (Chapman & Hall, 2004).
192 Patil, A., Huard, D. & Fonnesbeck, C. J. PyMC: Bayesian stochastic modelling in Python. *J. Stat. Softw.* **35**, 1-1, (2010).
193 Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. *Markov chain Monte Carlo in practice*. (Chapman and Hall/CRC, 1996).
194 Gelfand, A. E. & Smith, A. F. M. Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.*, 398-409, (1990).
195 Geweke, J. & Minneapolis, F. R. B. o. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. (Federal Reserve Bank of Minneapolis, Research Department, 1991).
196 WHO. *Western Pacific Region Office (WPRO) website*, <http://www.wpro.who.int/emerging_diseases/annual.dengue.data.wpr/en/index.html> (2012).
197 WHO. *Situation update of dengue in the SEA Region, 2010*, <http://www.searo.who.int/LinkFiles/Dengue_Dengue_update_SEA_2010.pdf> (2010).
198 WHO. *Eastern Mediterranean Regional Office: Weekly Epidemiological Monitor*, <http://www.emro.who.int/surveillance-forecasting-response/meeting-reports/> (2012).
199 WHO. *DengueNet data query*, <http://apps.who.int/globalatlas/DataQuery/default.asp> (2011).
200 WHO. *Dengue haemorrhagic fever: Diagnosis, treatment, prevention and control. 2nd edition*, <http://www.who.int/csr/resources/publications/dengue/Denguepublication/en/> (1997).
201 Rigau-Perez, J. G. Severe dengue: the need for new case definitions. *Lancet Infect. Dis.* **6**, 297-302, (2006).
202 Ng, C. F. S., Lum, L. C. S., Ismail, N. A., Tan, L. H. & Tan, C. P. L. Clinicians' diagnostic practice of dengue infections. *J. Clin. Virol.* **40**, 202-206, (2007).
203 Hay, S. I. *et al.* Estimating the global clinical burden of Plasmodium falciparum malaria in 2007. *PLoS Med.* **7**, e1000290, (2010).
204 Yoon, I. K. *et al.* Under-recognized mildly symptomatic viremic dengue virus infections in rural Thai schools and villages. *J. Infect. Dis.*, (2012).
205 Endy, T. P. *et al.* Epidemiology of inapparent and symptomatic acute dengue virus infection: a prospective study of primary school children in Kamphaeng Phet, Thailand. *Am. J. Epidemiol.* **156**, 40-51, (2002).
206 Yew, Y. W. *et al.* Seroepidemiology of dengue virus infection among adults in Singapore. *Ann. Acad. Med. Singapore* **38**, 667-675, (2009).

207 McBride, W. J., Mullner, H., LaBrooy, J. T. & Wronski, I. The 1993 dengue 2 epidemic in Charters Towers, North Queensland: clinical features and public health impact. *Epidemiol. Infect.* **121**, 151-156, (1998).
208 Jacobs, J., Fernandez, E. A., Merizalde, B., Avila-Montes, G. A. & Crothers, D. The use of homeopathic combination remedy for dengue fever symptoms: a pilot RCT in Honduras. *Homeopathy* **96**, 22-26, (2007).
209 Kantachuvessiri, A. Dengue hemorrhagic fever in Thai society. *Southeast Asian J. Trop. Med. Public Health* **33**, 56-62, (2002).
210 Chaudhuri, M. What can India do about dengue fever? *BMJ* **346**, (2013).
211 Endy, T. P., Yoon, I. K. & Mammen, M. P. Prospective cohort studies of dengue viral transmission and severity of disease. *Curr. Top. Microbiol. Immunol.* **338**, 1-13, (2010).
212 Graham, R. R. *et al.* A prospective seroepidemiologic study on dengue in children four to nine years of age in Yogyakarta, Indonesia I. studies in 1995-1996. *Am. J. Trop. Med. Hyg.* **61**, 412-419, (1999).
213 Porter, K. R. *et al.* Epidemiology of dengue and dengue hemorrhagic fever in a cohort of adults living in Bandung, west Java, Indonesia. *Am. J. Trop. Med. Hyg.* **72**, 60-66, (2005).
214 Thein, S. *et al.* Risk factors in dengue shock syndrome. *Am. J. Trop. Med. Hyg.* **56**, 566-572, (1997).
215 Tien, N. T. *et al.* A prospective cohort study of dengue infection in schoolchildren in Long Xuyen, Viet Nam. *Trans. R. Soc. Trop. Med. Hyg.* **104**, 592-600, (2010).
216 Tuntaprasart, W. *et al.* Seroepidemiological survey among schoolchildren during the 2000-2001 dengue outbreak of Ratchaburi Province, Thailand. *Southeast Asian J. Trop. Med. Public Health* **34**, 564-568, (2003).
217 Vong, S. *et al.* Dengue incidence in urban and rural Cambodia: results from population-based active fever surveillance, 2006-2008. *PLoS Negl. Trop. Dis.* **4**, e903, (2010).
218 Sirivichayakul, C. *et al.* Dengue infection in children in Ratchaburi, Thailand: a cohort study. II. clinical manifestations. *PLoS Negl. Trop. Dis.* **6**, e1520, (2012).
219 Dinh The, T. *et al.* Clinical features of dengue in a large vietnamese cohort: intrinsically lower platelet counts and greater risk for bleeding in adults than children. *PLoS Negl. Trop. Dis.* **6**, e1679, (2012).
220 Blaylock, J. M. *et al.* The seroprevalence and seroincidence of dengue virus infection in western Kenya. *Travel Med. Infect. Dis.* **9**, 246-248, (2011).
221 Bandyopadhyay, S., Lum, L. C. S. & Kroeger, A. Classifying dengue: a review of the difficulties in using the WHO case classification for dengue haemorrhagic fever. *Trop. Med. Int. Health* **11**, 1238-1255, (2006).
222 Deen, J. L. *et al.* The WHO dengue classification and case definitions: time for a reassessment. *Lancet* **368**, 170-173, (2006).
223 Srikiatkhachorn, A. *et al.* Dengue hemorrhagic fever: the sensitivity and specificity of the world health organization definition for identification of severe cases of dengue in Thailand, 1994-2005. *Clin. Infect. Dis.* **50**, 1135-1143, (2010).
224 Gupta, P. *et al.* Assessment of World Health Organization definition of dengue hemorrhagic fever in North India. *J. Infect. Dev. Ctries.* **4**, 150-155, (2010).
225 Kalayanarooj, S. Dengue classification: current WHO vs. the newly suggested classification for better clinical application? *J. Med. Assoc. Thai.* **94 Suppl 3**, S74-84, (2011).
226 Deparis, X., Murgue, B., Roche, C., Cassar, O. & Chungue, E. Changing clinical and biological manifestations of dengue during the dengue-2 epidemic in French

Polynesia in 1996/97-description and analysis in a prospective study. *Trop. Med. Int. Health* **3**, 859-865, (1998).

227 Diaz, F. A., Martinez, R. A. & Villar, L. A. Criterios clínicos para diagnosticar el dengue en los primeros días de enfermedad. *Biomedica* **26**, 22-30, (2006).

228 Dietz, V. J. *et al.* Epidemic dengue 1 in Brazil, 1986: evaluation of a clinically based dengue surveillance system. *Am. J. Epidemiol.* **131**, 693-701, (1990).

229 Kalayanarooj, S., Nimmannitya, S., Suntayakorn, S., Vaughn, D.W., Nisalak, A., Green, S., Chansiriwongs, V., Rothman, A., Ennis, F.A. Can doctors make an accurate diagnosis of dengue infections at an early stage. *Dengue Bull.* **23**, 1-9, (1999).

230 Kalayanarooj, S. *et al.* Early clinical and laboratory indicators of acute dengue illness. *J. Infect. Dis.* **176**, 313-321, (1997).

231 Lima, V. L. *et al.* Dengue: inquerito sorologico pos-epidemico em zona urbana do Estado de Sao Paulo (Brasil). *Rev. Saude Publica* **33**, 566-574, (1999).

232 Premaratna, R. *et al.* A clinical guide for early detection of dengue fever and timing of investigations to detect patients likely to develop complications. *Trans. R. Soc. Trop. Med. Hyg.* **103**, 127-131, (2009).

233 Rodrigues, E. M. *et al.* Epidemiologia da infeccao pela dengue em Ribeirao Preto, SP, Brasil. *Rev. Saude Publica* **36**, 160-165, (2002).

234 Klaucke, D. N. in *Principles and Practice of Public Health Surveillance* (ed S.M. Teutsch, Churchill, R.E.) 158-174 (Oxford University Press, 1994).

235 Dengue Surveillance in the Americas in *Accelerating Progress in Dengue Control.* (ed M.E. Beatty) Available at: http://www.denguevaccines.org/sites/default/files/Dengue Surveillance in the Americas_Mexico City.pdf (Americas Dengue Prevention Board, 17-19 January, 2008).