Supplemental Information Appendix

Methods

Occurrence data collection

A two year data collation process was conducted as part of the HumBug Project to build on the existing MAP database of global occurrence data for the Asian DVS of human malaria. Using the MAP data abstraction protocol [25, 70], additional searches of the published literature (PubMed, Web of Science and Scopus) were conducted extending the occurrence data from 2010 to 2016 and where possible, disaggregated sibling species data from the existing data across the full time period (1985 – 2016). Only those records with reliable reported identification methods (i.e. molecular methods distinguishing siblings) were included.

For the modelling presented here, we used *An. stephensi* occurrence data drawn from this updated dataset, supplemented with the most recent reports of *An. stephensi,* including those in the Horn of Africa and Sri Lanka.

Background data

Most species distribution modelling methods require absence/pseudo-absence or background data (true absence data is rarely available) [38, 71]. Ideally these data should reflect the same sampling bias as is found in the occurrence data (for example, occurrence data will be biased towards more accessible locations where surveyors can conduct their sampling) [72]. To do this, many background datasets used in mosquito vector mapping are based on sites where mosquito sampling has occurred, but where the target species was not recorded (e.g. [73]). These background points cannot be called 'true absences' as it may be that the target species does exist at the location but was just not found in the reported survey. Our initial set of background points included presences for anopheline species (excluding *An. stephensi*) across Asia and within the Horn of Africa, and records of *Aedes* and *Culex* species (some of these species have larval site behaviours similar to *An. stephensi*). To try and lessen the likelihood of a background point in reality being a presence point, but prevent the background points being selected from locations too far from the presence points (and therefore not representing the environmental space sampled for *An. stephensi*), we drew a polygon encompassing the occurrence data and established the distance between the central point within this polygon to the occurrence point the furthest distance away. This distance defined the buffer radius drawn around each occurrence point and only those background points (selected from within our existing dataset) within these buffers were sampled to provide the background data for the models. A buffer of 0.5 degrees was drawn around all the background points and a second buffer of 0.2 degrees was drawn around the presence points (to prevent background data being selected too close to the presence data). These three buffer zones in combination reflect the sampling bias implicit in the presence data.

As model accuracy can vary according to the selection of these points, we made 10 random selections across the defined background zone. The total number of pseudo-absence points (3762) per selection equates to 1% of the total number of pixels available in the background area.

Environmental Covariates

An initial suite of 19 covariates with varying resolutions were resampled to 0.1 degree (~11km) resolution to match the precision of our occurrence data (10km). To limit multi-collinearity we used principal component analyses (PCA), keeping only those variables with Spearman correlation coefficient <0.7 and those with specific relevance to mosquito biology, giving a final set of seven predictor variables (Table S2).

These seven covariate surfaces were downloaded from the source (See Table S2) and underwent minor pre-processing (coded in R [28, 29]) prior to use in *biomod2*. This included insuring all surfaces were in the same format (i.e. had the same projection) and any no data cells designated as '-9999' were assigned as 'NA' (the accepted notation in R). The rasters were stacked and masked to only include the relevant continents (Asia, Oceania and Africa) for *An. stephensi*.

The final set of seven covariates were: annual mean temperature and seasonal precipitation (Deblauwe et al [74] developed from MODIS [75] and CHIRPS [76] respectively), human population density (WorldPop [77, 78]), EVI (Enhanced Vegetation Index, corrected mean across each month and year for the period 2000-2014, MODIS satellite imagery [75]), Tasselled Cap Wetness (TCW) (Tasselled Cap Wetness – a measure of surface water, corrected mean across each month of the year for the period 2000-2014, MODIS satellite imagery [75]), Irrigation – Global Map of irrigated areas (GMIA) FAO, United Nations [79] and Crop mosaic (Proportional cover of land with a mosaic of croplands, forest, shrublands and grasslands, International Geosphere and Biosphere Programme (GBP) land cover classification within the MODIS dataset [80]).

Species Distribution Modelling

There are an increasing number of algorithms that can be applied to model species ecological niches and distributions, all with their own strengths and weaknesses [81, 82]. Different algorithms will have different conceptual approaches and different accuracy depending on the case study. The quality and quantity of the input data will also have a huge impact on the reliability, confidence and suitability outputs. To leverage the strengths and mitigate the weaknesses of such diverse modelling methods [83], we used *biomod2*: "… *a platform for ensemble forecasting of species distributions, enabling the explicit treatment of model uncertainties and the examination of species-environment relationships.'* [30]. This platform allows comparison between multiple species distribution models and those that perform well are then combined to create the final ensemble map. Our maps are composed from 500 model runs, using five species distribution modelling algorithms (Maximum Entropy (MAXENT) [32, 33], random forest [34] , generalised boosted regression [35], generalised additive models [36], and multiple adaptive regression splines (MARS) [37]).

The models were calibrated and evaluated over 10 data-splitting (80-20%) runs resulting in 500 model outputs. Two evaluations were used, the True Skill Statistic (TSS) and the area under the receiver operating curve (ROC), both based on a confusion matrix (a matrix that summarises the number of true and false predictions for presence and absence made by the model using a held back subset of data). The TSS is considered an improved substitute for the kappa value, a widely used metric but one that may be biased by the prevalence of the modelled species. It generates a confusion matrix by comparing a hypothetical set of perfect predictions to the number of correct predictions minus those attributed to random guessing. The output is a range between -1 to +1 where +1 equates to a perfectly agreed model whereas anything at zero or below indicate a model performance no better than random [84]. The ROC assessment plots the true prediction rate against the false prediction rate of multiple threshold data splits and the area under the subsequent curve provides a relative value for model performance [85].

The resulting consensus maps, indicating the habitat suitability for *An. stephensi* across Asia and into Africa, is an ensemble of the consensus mean of predictions of all 500 model runs that had a TSS and ROC evaluation score of > 0.6 and > 0.8 respectively. The outputs are weighted by the model's respective evaluation scores (i.e. TSS) before being combined.

All modelling was conducted in R studio (version 1.2.1335) [29] using R (version 3.6.1) [28].

Coefficient of Variation (CV) Map

Anopheles stephensi appears to be rapidly expanding its geographic range. Most SDMs assume some level of equilibrium between the species and its environment [38] [39] and species that are invading and expanding into new locations are not stable. Although we have included all available occurrence data for this species in Africa alongside suitable background points, it is important to note that we are still extrapolating into novel environments. Using an ensemble model methodology addresses this instability to some extent [38, 40, 41], but we are still attempting to predict occurrence in environments that are not represented by the sampled locations. Therefore we need to highlight where the modelled outputs are less robust.

We therefore accompany each of our ensemble models with a coefficient of variation (CV) map. This is a simple evaluation that maps the relative standard deviation (RSD – the ratio of the SD to the mean) across the 500 final models calculated on a per pixel basis. A high score therefore indicates a greater distribution around the mean and therefore a lower confidence in the mean ensemble value.

Model-selected influential variables

To establish the influence of each covariate, correlation scores were calculated on a model by model and covariate by covariate basis. Correlation values were calculated for each of the 500 models that made up the ensemble map, comparing the final model with an adjusted version where the values of the covariate in question had been shuffled/randomised. The correlation between the two models indicates the importance of that covariate. A low value (min = 0) indicates minimal influence of that covariate in the final model and a high value (max = 1) denotes a high influence on the final model [31]. These values were ranked to show their relative influence in the final models.

Population at risk

Risk is a product of predicted impact and likelihood. Three components of the likelihood of *An. stephensi* establishing and transmitting the malaria parasite in the future were calculated for the African cities with greater than one million inhabitants [42]. Using QGIS [43] a rectangle was drawn around the locations known to have *An. stephensi*, the quartile values were then calculated for the set of predictions within this rectangle and these were used to classify each of the African cities (using the maximum suitability value from the pixels within that city). Class 4 (below the lower quartile value, indicating lowest suitability) was not assigned to any listed city. Class 3 and 2 indicate increasing increments of suitability, with class 1 indicating the highest suitability (i.e. predictions greater than the upper quartile value). The distance from a confirmed *An. stephensi* record (occurrence data from this study) was calculated in QGIS using UN-sourced city coordinates. These coordinates were also used to evaluate the distance from the combined *P. falciparum* and *P. vivax* endemic zone, which was defined using the transmission limits data available from the Malaria Atlas Project [17].

Supplemental Tables

Table S1: The environmental covariates ranked by importance as selected by the inclusive and exclusive ensemble models.

	Exclusive map – WITHOUT African data		**Inclusive map – WITH African data**	
Model evaluation	TSS: 0.897, ROC: 0.985		TSS: 0.907, ROC: 0.987	
1	Annual Mean Temperature	0.459	Annual Mean Temperature	0.461
2	Human population density	0.325	Human population density	0.370
3	Precipitation Seasonality	0.171	Enhanced vegetation Index (EVI)	0.174
4	Enhanced vegetation Index (EVI)	0.161	Precipitation Seasonality	0.161
5	Irrigation	0.155	Tasselled cap wetness	0.134
6	Tasselled cap wetness	0.110	Irrigation	0.130
7	Cropland-Natural Vegetation Mosaic	0.011	Cropland-Natural Vegetation Mosaic	0.010

Table S2 Predictor variables. Descriptions of each potential explanatory variable used in the ensemble model. If the data layer was obtained from an online repository, the URL and date accessed are given. If the data layer has a citation then this is given.

Short name	Description	URL	Date accessed	Citation
Annual mean temperature	Annual mean temperature derived from MOD11C3 v. 5.0 (and published in [74]. Original data: MOD11C3 (MODIS sensor, 2001-2013) land surface temperature	https://vdeblauwe.wordpress.com/download/	2018	[74, 75]
Seasonal precipitation	Seasonal precipitation derived from CHIRPS v. 2.0 (and published in [74]. Original data: CHIRPS (1981-2013) precipitation dataset.	https://vdeblauwe.wordpress.com/download/	2018	[74, 76]
Tasselled Cap Wetness (TCW) - Surface wetness mean	Mean values for a measure of surface moisture (TCW, variation in the vigour of green vegetation)	https://lpdaac.usgs.gov/products/mcd43d6*2-4*v006/	2018	[86]
Cropland-natural vegetation percentage (Class 14)	Proportion of the pixel area covered by a mosaic of annual crops and natural vegetation (mosaic of cropland, forest, shrubland or grassland)	https://modis.gsfc.nasa.gov/data/dataprod/mod12.php	2018	[75]
Enhanced Vegetation index (EVI) mean	Mean enhanced vegetation index is a measure of greenness reflectance of the land surface	https://lpdaac.usgs.gov/products/mcd43d6*2-4*v006/	2018	[87]
Irrigation	Global Map of irrigated areas (GMIA)	http://www.fao.org/aquastat/en/geospatial-information/global-maps-irrigated-areas	2018	[79]
Population density	WorldPop, human population size (No. persons/pixel)	https://www.worldpop.org/geodata/listing?id=17	n/a	[78]

Table S3: The populations at risk if *An. stephensi* were to establish in the urban cities of Africa.

City	Country	Population	Distance from *An. stephensi* records (km)	Distance from malaria endemic zone (km)	Habitat suitability class
Djibouti City	Djibouti	<1M	0	0	1
Addis Ababa	Ethiopia	3,725,000	160	[1]0	2
Asmara	Eritrea	<1M	450	0	1
Muqdisho (Mogadishu)	Somalia	1,890,000	480	0	1
Al-Khartum (Khartoum)	Sudan	6,150,000	940	0	1
Nairobi	Kenya	5,950,000	1100	0	1
Mombasa	Kenya	1,240,000	1190	0	1
Kampala	Uganda	3,400,000	1280	0	2
Dar es Salaam	Tanzania	6,150,000	1490	0	1
Kigali	Rwanda	1,140,000	1650	0	2
Kisangani	DRC	1,120,000	1900	0	3
Al-Qahirah (Cairo)	Egypt	20,500,000	2260	1490	1
Bangui	CAR	1,160,000	2440	0	2
Lilongwe	Malawi	1,020,000	2450	0	1
Mbuji-Mayi	DRC	2,000,000	2490	0	2
Kananga	DRC	1,190,000	2570	0	3
Lubumbashi	DRC	2,125,000	2640	0	2
N'Djaména	Chad	1,360,000	2720	0	1
Tananarive (Antanarivo)	Madagascar	2,450,000	[2]2780	0	1
Lusaka	Zambia	2,575,000	2900	0	1
Maiduguri	Nigeria	1,170,000	2930	0	1
Harare	Zimbabwe	2,050,000	3080	0	2
Brazzaville	Congo (Rep.)	2,175,000	3130	0	1
Kinshasa	DRC	12,000,000	3130	0	1
Yaoundé	Cameroon	3,300,000	3220	0	2
Douala	Cameroon	3,250,000	3410	0	1
Jos	Nigeria	1,000,000	3410	0	2
Kano	Nigeria	4,550,000	3430	0	1
Pointe-Noire	Congo (Rep.)	1,130,000	3490	0	1
Kaduna	Nigeria	1,830,000	3560	0	1
Abuja	Nigeria	3,375,000	3570	0	1
Luanda	Angola	7,900,000	3580	0	1
Tarabulus (Tripoli)	Libya	1,210,000	3640	1880	1
Aba	Nigeria	1,120,000	3650	0	2
Onitsha	Nigeria	1,170,000	3700	0	1
Port Harcourt	Nigeria	2,275,000	3700	0	2
Maputo	Mozambique	2,875,000	3730	0	1
Benin City	Nigeria	1,610,000	3820	0	1
Ilorin	Nigeria	1,040,000	3900	0	1
Johannesburg/	South Africa	13,700,000	3930	300	1

Pretoria					
Ibadan	Nigeria	3,475,000	3990	0	1
Sousse	Tunisia	1,120,000	4020	2110	2
Lagos	Nigeria	18,800,000	4060	0	1
Tunis	Tunisia	2,625,000	4120	2210	2
Niamey	Niger	1,260,000	4130	0	1
Cotonou	Benin	1,770,000	4170	0	1
Durban	South Africa	3,375,000	4180	150	3
Lomé	Togo	2,025,000	4300	0	1
Accra	Ghana	4,950,000	4470	0	1
Ouagadougou	Burkina Faso	2,525,000	4530	0	1
Kumasi	Ghana	2,950,000	4600	0	1
El Djazaïr (Algiers)	Algeria	3,900,000	4660	2160	2
Port Elizabeth	South Africa	1,330,000	4800	840	3
Abidjan	Ivory Coast	5,450,000	4890	0	1
Wahran (Oran)	Algeria	1,450,000	4910	2000	2
Cape Town	South Africa	4,175,000	5150	1220	2
Fès	Morocco	1,250,000	5220	1820	1
Bamako	Mali	3,375,000	5230	0	1
Tangier	Morocco	1,090,000	5350	2020	2
Rabat	Morocco	2,025,000	5380	1840	1
Marrakech	Morocco	1,100,000	5410	1550	1
Dar-el-Beida (Casablanca)	Morocco	4,425,000	5430	1770	1
Agadir	Morocco	1,160,000	5530	1400	1
Monrovia	Liberia	1,530,000	5610	0	2
Freetown	Sierra Leone	1,720,000	5840	0	1
Nouakchott	Mauritania	1,220,000	6060	0	2
Touba	Senegal	1,040,000	6060	0	2
Dakar	Senegal	3,600,000	6230	0	1

[1] Addis Ababa is within the area of the malaria endemic zone, but due to its altitude (~2300m) is considered non-endemic.

[2] This Euclidean distance includes a stretch that traverses the Mozambique Channel/Indian Ocean.

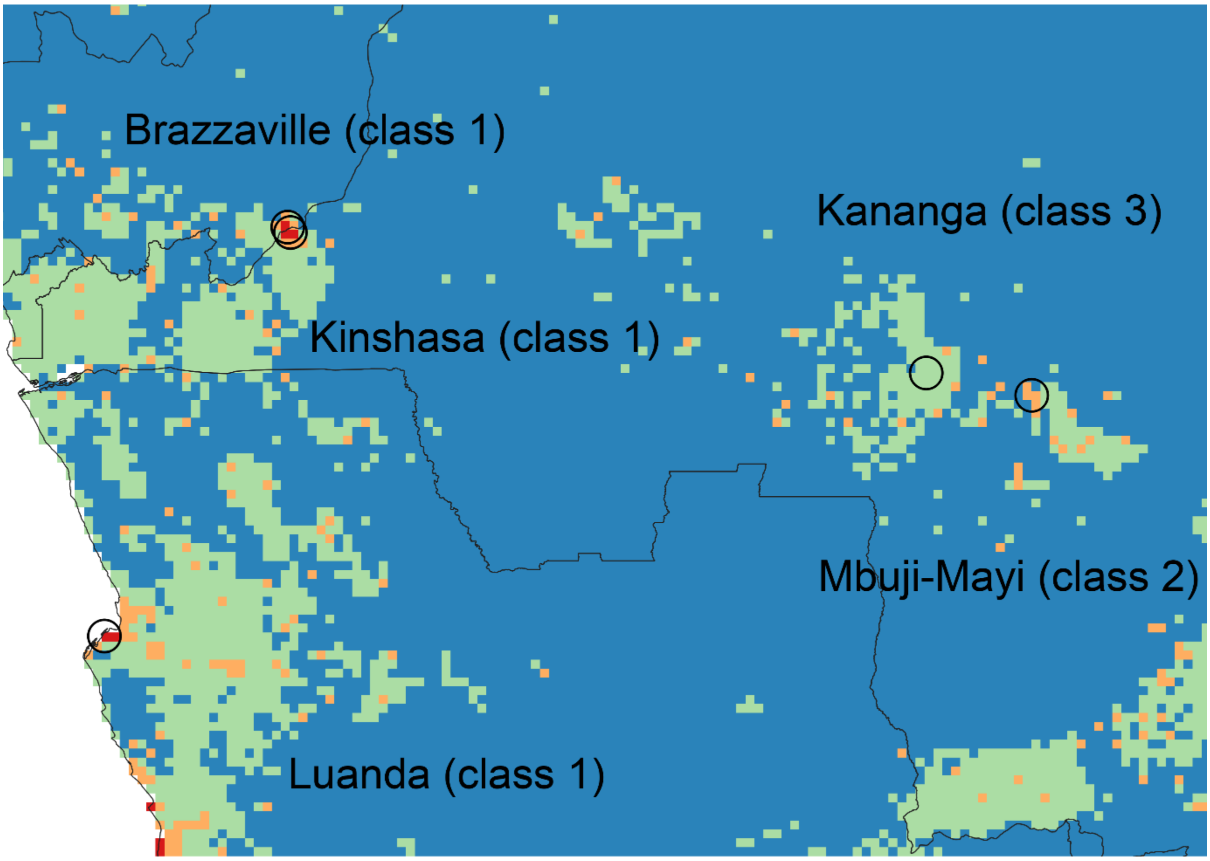

Figure S1: Populations at risk calculations: close up indicating examples of how the cities were classified showing examples of cities in high (Class 1), medium (Class 2) and low (Class 3) risk.

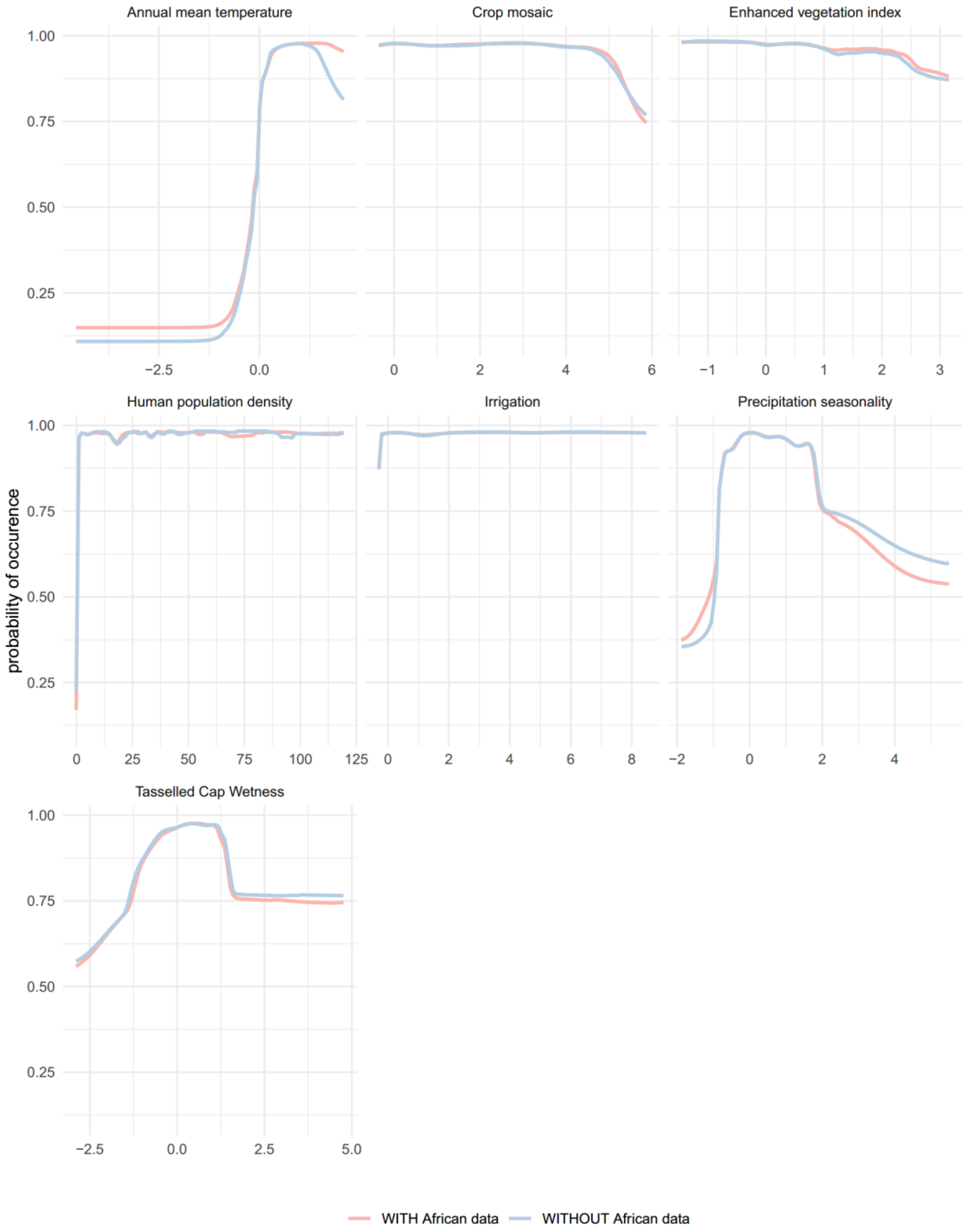

Figure S2: Response curves for the seven explanatory covariates derived from models run WITHOUT (exclusive) (n = 343, blue line) and WITH (inclusive) the African data (n = 358, red line). The annual mean temperature and human population density variables has the most influence in both of the final models.