**Seasonal Dynamics of an Emerging African Malaria Vector, *Anopheles stephensi*: Implications for Malaria Establishment and Control**

Charles Whittaker1\*, Peter Winskill1, Gina Cuomo-Dannenburg1, Patrick G.T. Walker1, Marianne Sinka2, Ashwani Kumar3, Azra Ghani1, Thomas Churcher1, Samir Bhatt1,4 & Arran Hamlet1

1MRC Centre for Global Infectious Disease Analysis & Abdul Latif Jameel Institute for Disease and Emergency Analytics, School of Public Health, Imperial College London, London, UK

2Department of Zoology, University of Oxford, Oxford, UK

3Vector Control Research Centre, Indira Nagar, Puducherry, India

4Section of Epidemiology, Department of Public Health, University of Copenhagen, Copenhagen, Denmark

**\*Corresponding Author:** Charles Whittaker, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College, London W2 1PG, United Kingdom. Email: [charles.whittaker16@imperial.ac.uk](mailto:charles.whittaker16@imperial.ac.uk)

**Contents**

**Supplementary Information 1:** Description of Systematic Review, Data Extraction and Initial Processing

**Supplementary Information 2:** Description of Statistical Methodologies Utilised

**Supplementary Information 3:** Additional Figures and Results

**References**

**Outline of Document**

In this supplementary document we outline the methods and data used to explore and analyse the patterns and drivers of *Anopheles stephensi* population dynamics across South Asia and the Middle East. In **Supplementary Information 1**, we present an overview of the systematic search strategy used to collate the references containing the extracted and analysed data, as well as details about the initial pre-processing steps applied to said data. In **Supplementary Information 2**, we describe the statistical methodologies used to process and analyse this extracted data. The output of these analyses forms the basis for the results presented in the main text. Finally, in **Supplementary Information 3**, we present a number of additional figures and tables to support the work detailed in the main text.

**Supplementary Information 1: Description of Systematic Review: Data Extraction and Initial Pre-Processing**

**Systematic Review: Search Procedure and Record Screening**

We collated references from previously published systematic reviews of literature relating to *Anopheles stephens* (focusing on its presence/absence across a wide geographical range1 and its seasonal dynamics in India2 respectively), and updated these previous searches (both conducted in 2017) by searching *Web of Science* and *PubMed* databases from January 2017 to September 2020 for further relevant references containing temporally disaggregated *Anopheles stephensi* catch data. Key words for this search were:

(((anophel\*) AND ((India) OR (BURMA) OR (MYANMAR) OR (BANGLADESH) OR (THAILAND) OR (ISLAMIC REPUBLIC OF IRAN) OR (ETHIOPIA) OR (DJIBOUTI) OR (SUDAN))) AND (("2017"[Date - Publication] : "3000"[Date - Publication])) OR ((anophel\*) AND ((Pakistan) OR (Iran) OR (Afghanistan)) AND (("1990"[Date - Publication] : "3000"[Date - Publication]))

with references for Pakistan, Iran and Afghanistan searched for over an extended time-period (i.e. date range of 1990-2020 rather than 2017-2020) to ensure completeness of the collated references, and fill in countries not included during previous reviews. Our searches identified a total of **X** records, which were screen according to the following Inclusion/Exclusion criteria:

**Inclusion Criteria:**

* Reference contains temporally disaggregated adult mosquito catch data at a temporal resolution of monthly or higher.
* The time-period spanned by the survey must be at least 10 months in duration and have caught at least a total of 25 *Anopheles stephensi* over the period for which catches were being carried out.

**Exclusion Criteria:**

* Mosquito catch data is not temporally disaggregated to a sufficient extent (e.g. catches were done yearly or seasonally rather than monthly).
* Mosquito catch data was collected as part of a trial assessing a vector control intervention (which would perturb the natural dynamics of the vector, rendering the data unrepresentative of the population dynamics in the absence of control).
* The reference only contained information on immature/larval mosquito life cycle stages rather than mature adults.
* The reference contained insufficient information to geolocate the area in which the study was conducted to at least the administrative unit 2 level.

Overall, a total of 32 references were collated containing 62 time-series from catch surveys carried out in distinct locations from across Afghanistan (n=2), Djibouti (n=1), India (n=31), Iran (n=17), Myanmar (n=5) and Pakistan (n=7). These were further supplemented with 2 references (from Pakistan and India respectively, yielding a total of 3 time-series) collated as part of a (currently unpublished) systematic review of the bionomics of secondary malaria (i.e., non-dominant) vectors across South Asia, yielding a total of 65 time-series from these 34 references. The next section describes in further detail about extraction and collation of the data associated with each study.

**Systematic Review: Data Extraction, Collation and Initial Processing**

**Entomological Data Extraction**

For each reference, we extracted all relevant entomological catch data provided that pertained specifically to *Anopheles stephensi*. Where data were presented in a table, data was copied directly from the table. Where the data were in a graph, data were extracted using the DataThiefTM software. This yielded a total of 65 time series of monthly mosquito catch data (no reference presented data at a finer temporal resolution), ranging in length from 10 – 60 months, with a mean time-period of 15.6 months and a median time-period of 12 months, a mean catch size of 758 and a median catch size of 289.

The primary focus of these analyses was to specifically characterise annual and seasonal patterns of variation in *Anopheles stephensi* abundance. Given this, and also that variations in time-series length are a factor known to affect their statistical properties3 (and therefore limit the comparability of the time series gathered and analysed here), all time series were standardised to be 12 months in length. For time series containing more than 12 time points (i.e. time series that spanned longer than a single year), we averaged the recorded catches for a given month. For time-series containing less than 12 months of data, this was not carried out. Where the study has been initiated in a month other than January, and concluded in a month other than December, the recorded counts were rearranged to yield a complete time series running from January to December.

The results presented in the collated references were frequently presented in the form standardised by sampling effort, such as Man-Hour Density (MHD). They do not therefore represent the total number of mosquitoes caught each month (required for the statistical framework utilised to characterise temporal properties) and therefore, where information on where information on sampling effort was present (e.g. number of hours spent sampling/catching *Anopheles stephensi*, number of households or cattlesheds searched, number of trap nights etc), we used this information to convert MHD back to the raw counts. In the small number of instances where there was variable sampling effort across the time series (which would bias the conversion away from the underlying population abundance), we conservatively used the lowest sampling effort recorded across the time series in the conversion. Together, this allowed us to produce an estimate of the number of mosquitoes sampled (a raw count, based on equal sampling effort across the time series). See **Supplementary Data: *Extracted Entomological Data***for more information about how each time-series was processed).

**Study Geolocation and Environmental Covariate Extraction**

For each study where geolocation was possible, we recorded the location at both the administrative unit 1 and 2 level, based on information provided in the reference. A number of the references identified in our review had previously been utilised as part of previous reviews1,2 – where this data was available, these descriptions of study location were used. For each location, we then extracted a suite of satellite-derived environmental covariates. These environmental covariates consist of raster layers spanning all of the countries studies had been conducted in (i.e. Afghanistan, Djibouti, India, Iran, Myanmar and Pakistan) at a 2.5 arc-minute (~5km by 5km, depending on the exact location and distance from the equator) spatial resolution. The covariates utilised here were initially selected from a set of X derived from the *BioClimatic* variables (a suite of biological relevant covariates defined from monthly rainfall and temperature satellite data4) as well as measures of landcover and urbanicity5, population density6,7 and enhanced vegetation index8,9. This provided a total of **X** covariates, many of which were highly correlated with one another. To reduce the degree of this multicollinearity, we generated a reduced subset of covariates using tools available in the *tidymodels* collection of R packages10 that aim to minimise the Spearman correlation coefficients between retained covariates, and also exclude covariates where there is minimal variation for that (scaled and standardised to have mean 0 and unit variance) covariate across the full dataset, leaving **X** covariates in total. See **Supplementary Figure X** for the correlation matrix of the retained variables.

In addition to the environmental covariates described above, for each of the administrative units a survey had been carried out in, we also collated daily rainfall estimates for the time-period the survey had been conducted in, using the *“The Climate Hazards Group Infrared Precipitation With Stations”* (CHIRPS) dataset11. These data were aggregated up to the same temporal resolution as the *Anopheles stephensi* catch data (i.e. monthly). These rainfall data were used to calculate the cross-correlation coefficient between mosquito catches and rainfall.

**Supplementary Table 1: The Complete Suite of Environmental Covariates Collated and Subsequently Reduced for Modelling and Prediction of Seasonal Population Dynamics**

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Variable** | **Temporal Resolution** | **Source** |
| 1 | BioClimatic - Annual Mean Temperature | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 2 | BioClimatic - Mean Diurnal Range | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 3 | BioClimatic - Isothermality | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 4 | BioClimatic - Temperature Seasonality | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 5 | BioClimatic - Max Temperature of Warmest Month | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 6 | BioClimatic - Min Temperature of Coldest Month | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 7 | BioClimatic - Temperature Annual Range | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 8 | BioClimatic - Mean Temperature of Wettest Quarter | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 9 | BioClimatic - Mean Temperature of Driest Quarter | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 10 | BioClimatic - Mean Temperature of Warmest Quarter | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 11 | BioClimatic - Mean Temperature of Coldest Quarter | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 12 | BioClimatic - Annual Precipitation | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 13 | BioClimatic - Precipitation of Wettest Month | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 14 | BioClimatic - Precipitation of Driest Month | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 15 | BioClimatic - Precipitation Seasonality | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 16 | BioClimatic - Precipitation of Wettest Quarter | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 17 | BioClimatic - Precipitation of Driest Quarter | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 18 | BioClimatic - Precipitation of Warmest Quarter | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 19 | BioClimatic - Precipitation of Coldest Quarter | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 20 | Population Density | **CHECK WITH ARRAN** | <http://www.worldpop.org.uk> |
| 21 | Enhanced Vegetation Index | **CHECK WITH ARRAN** | **CHECK WITH ARRAN** |
| 22 | Landcover | **CHECK WITH ARRAN** | **CHECK WITH ARRAN** |

**Note:** There are **X** covariates total here, as Landcover contains **X** distinct covariates (each describing the proportion of cover attributable to a particular landcover class in a given area).

**Note:** All WorldClim data is from Version 2 of the datasets.

**Supplementary Information 2: Description of Statistical Methodologies**

**Negative Binomial Gaussian Process – Fitting and Inference:**

In-line with previously work modelling the seasonal dynamics of different *Anopheline* mosquito species from across India2, we utilise a flexible Gaussian Process modelling framework to temporally interpolate between the monthly-catch datapoints and smooth the raw, noisy and overdispersed catch data. Gaussian processes specify a distribution over functions such that any finite set of function values have a joint Gaussian distribution12. The Gaussian process is entirely specified by its mean function:

and by its covariance function:

The covariance function is also known as the kernel and defines, based on the Euclidean distance between any two points, their covariance (and thus the covariance matrix of the Gaussian Process when all pairwise combinations of points are considered). Many different forms of the kernel are possible that each encode different prior information about how we expect two datapoints ( and in this instance) to be similar, and the distance over which we expect this similarity to persist. Given that mosquito population dynamics are typically characterised by repeating patterns occurring either seasonally or annually, a periodic kernel function was used to define the covariance between pairs of points:

where represents the period over which we would expect points to show similar dynamics (i.e. a period of twelve would imply we expect points separated by 12 months to be most similar), specifies the magnitude of the covariance, and  represents a lengthscale parameter further constraining the extent to which two values separated by a given time can co-vary.

Bayesian inference and fitting of normal Gaussian Processes typically follow this hierarchical formulation:

where represents a vector of hyperparameters involved in defining the kernel’s properties, is a distribution of functions from a zero-mean Gaussian Process with covariance function , (x) are function evaluations at times , and the observed data. We modify this structure to account for specific characteristics of the mosquito data being utilised – specifically that the data are integer counts, that mosquito catch data is rarely normally distributed and frequently displays high levels of overdispersion (a common property of biological systems generally). We therefore adapted the above framework to accommodate a Negative Binomial likelihood, leading to the following inferential framework:

where is used to reflect the fact that we use a log link between the observed counts and the underlying latent process reflecting the population dynamics, and represents the overdispersion parameter of the Negative Binomial distribution.

**Prior Specification**

Per previous work2, prior distributions for the estimated parameters were defined as follows:

Weakly informative priors were set on the scaling factor , the period, , and the overdispersion parameter, . The prior for the kernel period () was centred on (a value of the period that would represent annual variation being the dominant temporal modality) to reflect our prior belief that observed variation in mosquito abundance is likely to cycle annually. However, recognising that other temporal patterns of fluctuating abundance are possible, we placed a large standard deviation on to allow the model to accommodate instances of bimodality or periods operating across timescales longer than a year. We placed lower and upper bounds on at 4 and 18 months respectively, to avoid identifiability issues arising from the lack of data at temporal resolutions below and above these bounds.

**Model Fitting and Parameter Inference**

This Negative Binomial Gaussian Process were fitted using a Bayesian framework implemented in STAN, a probabilistic programming language for statistical inference written in C++ that employs the No-U-Turn sampler, a variant of the gradient-based Hamilton Monte Carlo algorithm for inference13. For each time-series, **X** chains of **X** iterations were run for purposes of model fitting and parameter inference. Half of each chain’s iterations were discarded as burn-in/the adaptive phase of the sampling, leaving a total of **X** iterations available for inference. Measures of MCMC convergence such as the Gelman-Rubin statistic were monitored in all cases and were all consistently < 1.02.

**Fitted Time Series Normalisation and Von Mises Distribution Fitting**

After having fitted and smoothed the mosquito catch time-series, we normalised each in the following way:

where is the proportion of the annual catch recorded at timepoint . This was done in order to establish comparability across the time series (which varied substantially in the absolute numbers of *Anopheles stephensi* caught). We then further characterised the periodic properties of these time series by fitting Von Mises distribution to the time-series. The Von-Mises distribution is a continuous probability distribution that exists on the circle, with range to . It is the circular analogue of the normal distribution (which exists on the line), with the probability density function for the angle given by:

where is the modified Bessel function of order 0, the parameter is a measure of location (analogous to the mean of the normal distribution, describing where on the circle the distribution is clustered around) and describes the concentration of density around (and thus its inverse is a measure of dispersion, analogous to for the normal distribution. We fitted two sets of Von Mises densities to the normalised time series, the first containing a single component:

and another with two-components, formulated as:

where represents the normalised mosquito count formulated as a random variable on the circle (i.e. ). Fitting was undertaken using the *optim* function in R, with the root mean squared error as the loss function. The outputs from this fitting were then included in the process generating aggregate summaries of the temporal properties of the time-series, a process described in further detail below.

**Time Series Characterisation and Analysis**

To characterise the temporal properties of each time-series, we calculated a series of summary statistics for each, drawing on previous work carried out exploring the empirical structure of time series12. In doing this, we can make explicit comparisons between time-series about key aspects of their temporal properties (e.g., the degree or timing of seasonality), and in doing so, identify time-series with similar statistical and temporal properties. These summary statistics were the following:

1. **Periodic Kernel Median:** Fitting the Negative Binomial Gaussian Process with a periodic kernel allowed inference of the period, , providing us with an estimate of the frequency of repeating patterns in the monthly abundance of mosquitoes. An estimate of was calculated for each fitted time series, with the median value of across the X HMC iterations for each time-series used here
2. **Kullback-Leibler Divergence:** Also known as the relative entropy, the Kullback-Liebler divergence represents a measure of how different one probability distribution is from a second probability distribution (where a value of 0 indicates that the two distributions are identical). It is specified in the following manner:

where is the average value of the normalised time series for month , and = 1/12 for . This operation therefore measures the deviation of a normalised time series from a uniform distribution, in doing so, informing about the extent to which a seasonal peak (or peaks) is present in the time series.

1. **Time Difference Between Vector Density Peak and Rainfall Peak Timings:** The time difference between the highest recorded vector density and the highest recorded rainfall for that year.
2. **Proportion of Points Greater Than 1.65x the Mean:** For each fitted, normalised time series, the proportion of points greater than 1.65x the time-series’ mean was calculated, informing the degree and width of any seasonal peaks.
3. **Number of Peaks:** Estimates of the parameters governing the fitted two component Von Mises distribution were used to infer the number of peaks in each time series. Specifically, and in-keeping with previous work2, a time series was deemed to possess one peak if the value of the Von Mises component weighting was either < 0.3 or > 0.7 and the difference in means was < or > , indicating that the majority of the density could be attributed to one of the two components, and that the two means identified during the fitting were temporally close to one another. Otherwise, a time series was judged to possess two peaks.
4. **Von Mises 1 Component Mean:** If a 1 component Von Mises distribution was preferred, then the Von Mises mean corresponding to the maximum likelihood predicted value was used. If the 2 component Von Mises distribution was preferred, the value for this operation for that particular time series is set to -5.
5. **Von Mises Two Component Weight:** Estimates of the weight parameter governing the two component Von Mises distribution were also used to infer the bimodality of the time series. The weight specifies the proportion of each component that is used to fit the time series and thus a very high (or very low weight) indicates the dominance of a single component and the comparatively small contribution of the other.
6. **Maximum Percentage of Total Annual Catch In Any 4 Month Period:** In-keeping with previous, operationally aligned estimates of malaria seasonality14, we calculated using a sliding 3-month window the maximum percentage of the total annual catch that was caught in any 3 month period.

**Principal Components Analysis and Clustering**

PCA is a statistical procedure that utilises an orthogonal transformation to convert a set of correlated variables (in this case the outputs of the 7 mathematical operations described above for each of the time series) into a set of linearly uncorrelated variables (known as the “principal components”). In doing so, this allows us to summarise this set of variables with a smaller number of representative variables that together explain the majority of the variability in the variables. Reducing the dimensionality of the dataset in this way facilitates visualisation of time series properties (as defined by the mathematical operations) as well as clustering of the time series into groups which share similar properties (clustering algorithms typically perform poorly in high dimensional settings, necessitating the use of PCA as described here). Clustering was then undertaken using the k-means clustering algorithm, using the 4 first PCA components that together described **X**% of the total variation present in the data.

**Random Forest Modelling and Prediction of Seasonality**

Random Forests are a machine learning, ensemble-based method that work by constructing a collection of decision trees that together explain the results (where results are either a continuous outcome variable in the regression context, or a binary indicator in the classification context)15. The outputs of these decision trees are subsequently aggregated in a statistically principled and coherent way to produce a “forest” (or ensemble) of trees that together produce predictions for comparison with data. They have previous been shown to provide significant improvements in accuracy over traditional linear regression based approaches, particularly in contexts where non-linear relationships or interactions between covariates are likely present and to be relevant to prediction of an outcome16. For the analyses described below, random forest was carried out using the Ranger R package17, implemented within the *tidymodels* framework10.

We used a Random Forest based approach to predict whether an *Anopheles stephensi* time-series had a single seasonal peak (as defined in the above section ***“Time Series Characterisation and Analysis”***) or not (with this latter category including those with bimodal timeseries or those with more perennial patterns of abundance and less pronounced seasonal fluctuations). Out-of-sample predictive accuracy was assessed using 6-fold cross-validation (CV), and was used to optimise the hyperparameters associated with the Random Forest method algorithm. Random Forest models were fitted to the training dataset (i.e. the full dataset minus one of the CV folds) and then model accuracy assessed on the remaining fold of data not included in model training, with model accuracy assessed by calculating the area under the curve (AUC). The Random Forest hyperparameters providing the best out-of-sample AUC were then selected and a final Random Forest model then fitted on the full set of data available. Predictive accuracy (assessed via AUC) was then calculated for the entire dataset by using out-of-bag predictions for each sample i.e. predictions on each training sample using only the trees that did not have that training sample in their bootstrap sample. We also calculated both permutation variable importance and generated partial dependency plots18 for each model to assess the contribution of specific, individual environmental covariates to whether a time-series had a single seasonal peak or not. Together these methods allow evaluation of the importance of each included covariate to model predictive accuracy, and in turn, allows us to “rank” covariates according to their contribution to the predictive performance of the model. This entire process was repeated 25 times in order to average over the stochasticity and variation inherent in the Random Forest fitting process.

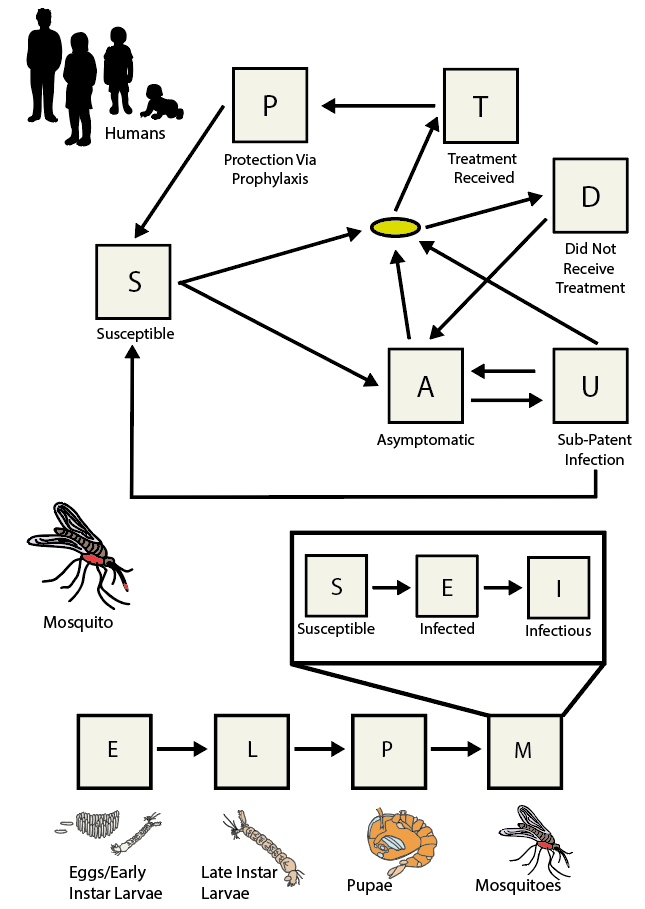
**Modelling of Malaria Transmission and the Impact of *Anopheles stephensi***

We integrated the temporal profiles of *Anopheles stephensi* abundance into a deterministic version of a well-established compartmental model of *Plasmodium falciparum* malaria transmission and disease19–21 to explore the implications of the vector’s establishment and seasonality on the dynamics of malaria transmission, with a particular focus on areas where malaria transmission is currently absent or only minimally present. What follows is a description of the mathematical modelling framework in general terms, followed by specific details about how exactly this framework was used to model malaria transmission underpinned by *Anopheles stephensi* establishing in settings where malaria is currently absent or only minimally present.

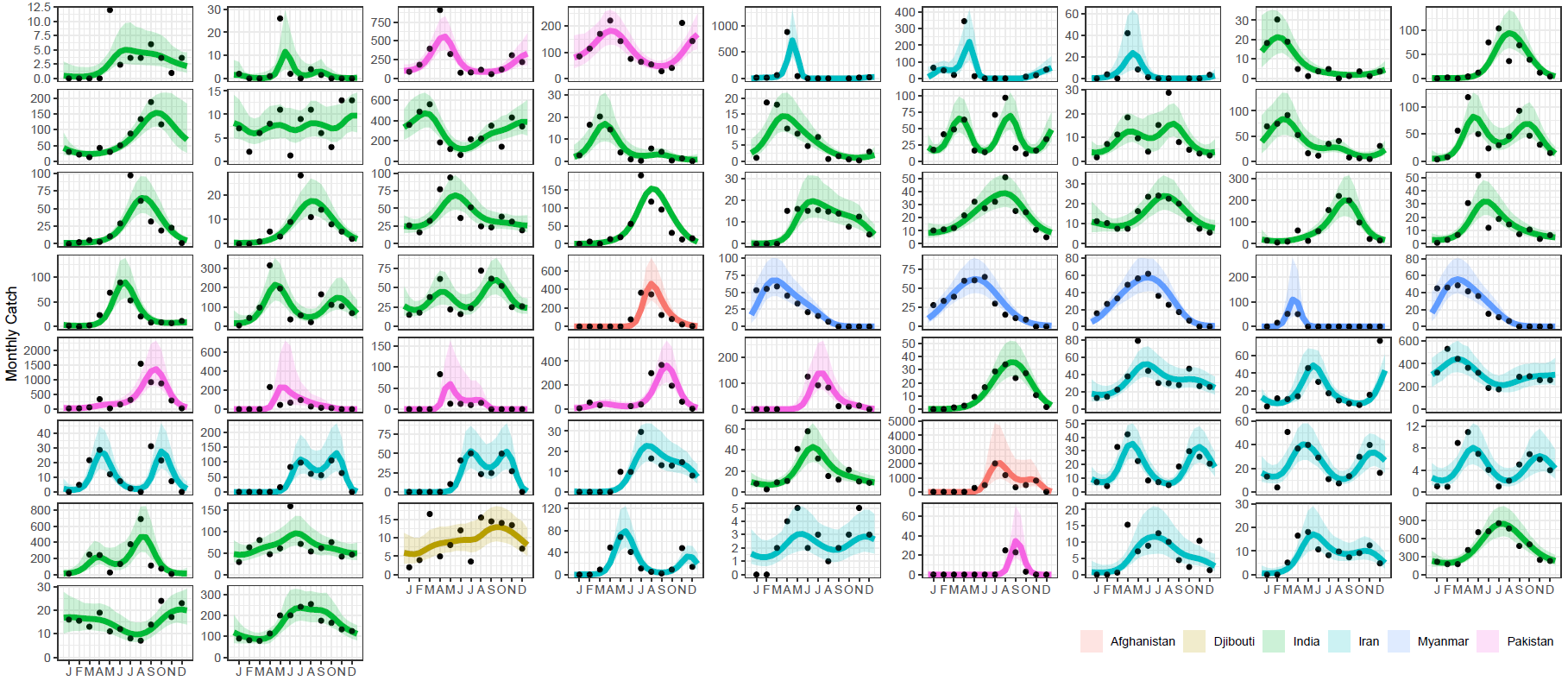
The deterministic malaria model used here considers both human and mosquito populations. Humans begin as Susceptible (S), and upon infection (at a rate which is dependent on the force of experience they experience), progress to either Asymptomatic (A) or clinical disease, with the comparative probability of these two outcomes depending on the degree of acquired natural immunity due to previous exposure to the parasite. If an individual progresses to clinical disease, they enter either a Treated (T) or Clinical Disease (D) state that depends on the probability of receiving treatment. For those treated, individuals progress through a period of prophylactic protection following treatment (P), and then return to the susceptible compartment. For those developing clinical disease, they remain symptomatic for the duration of the disease, before moving to an asymptomatic state (A, detectable by light microscopy), before subsequently moving to a submicroscopically infected state (U, not detectable by light microscopy). Individuals who are currently asymptomatically infected (including individuals in both the A and U states) can be reinfected and develop clinical disease once again – if this does not occur, they subsequently clear the infection and return to the susceptible state.

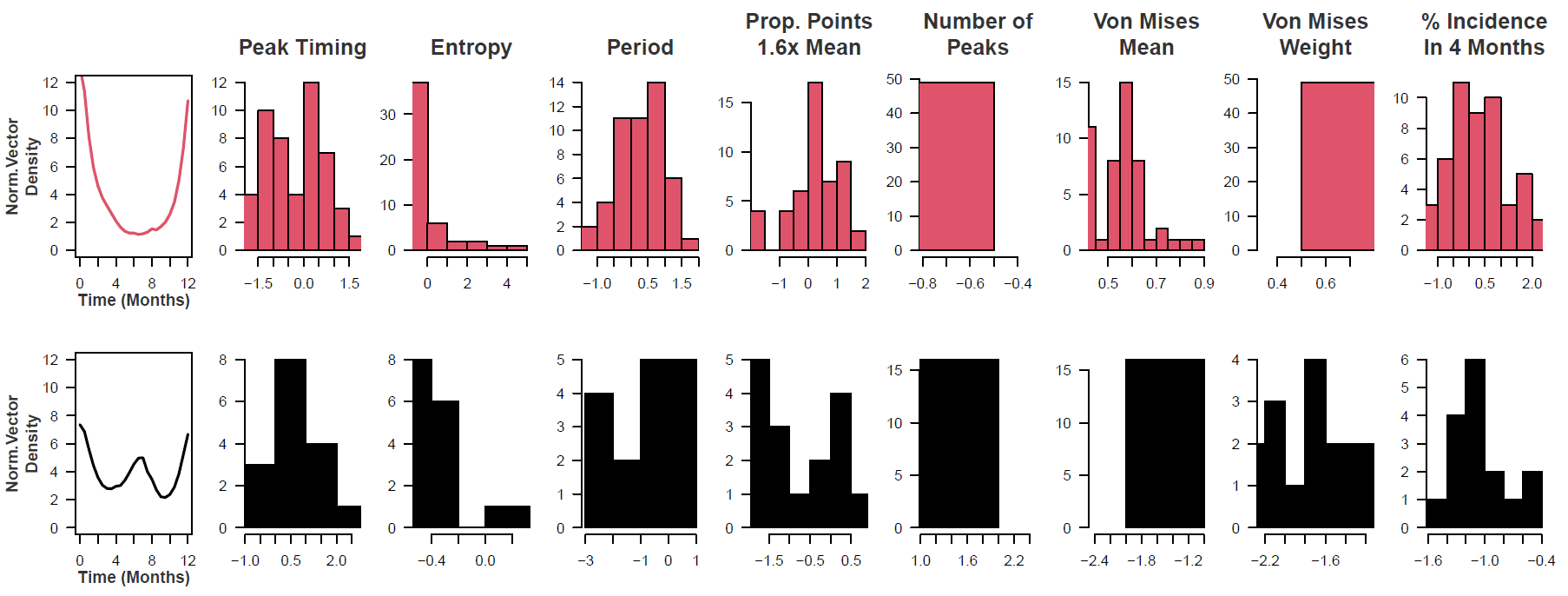
Adult mosquito populations and their preceding juvenile stages are also explicitly modelled. Immature mosquitoes start off as larvae, divided into early and late stage (Es and Ls respectively) which then mature into pupae (P) before eventually maturing into adult mosquitoes. Adult mosquitoes are further stratified according to infection with *Plasmodium falciparum* status – they begin as susceptible (Sm) and upon infection, progress to an exposed (but un-infectious, Em) state, and then onto the infectious state (Im) following the extrinsic incubation period (EIP). Mosquitoes are infected through exposure to humans currently possessing transmissible infections i.e. the treated (T), clinical disease (D), asymptomatic (A) and submicroscopic (U) infection states.

Seasonality in mosquito abundance is incorporated through a flexible, time-varying carrying capacity that in broad terms describes temporal variation in the ability of the local hydrological environment to support mosquito breeding. The value of this carrying capacity relative to the size of the mosquito population influences the mortality of early and late-stage larvae, which as previous modelling work has shown, enables the model to accurately and adequately capture temporal fluctuations in mosquito abundace22. We integrate each of the seasonal profiles of *Anopheles stephensi* density implied by the corresponding time-series of catch data into the model, matching the carrying capacity to the empirically observed temporal variation in *Anopheles stephensi* abundance. To simulate the invasion of *Anopheles stephensi* in setting where it (and malaria transmission more generally) is currently absent, we layer an additional time-varying component of vector density on top of these seasonal variations. Starting from 0 vector density (simulating an absence of malaria and of *Anopheles stephensi*), we increase the annual average vector density in a sigmoidal fashion over X years – seasonal fluctuations in density are then applied on top of this annual average density (see Supplementary Figures X and X for more detail on the seasonal and annual changes in vector density used in the model simulations). Estimates of the bionomic properties of *Anopheles stephensi* (specifically the mosquito’s daily mortality, degree of anthropophagy, degree of endophily and the proportion of bites taken on individuals indoors and/or in bed) were taken from previous work that reviewed the properties23.

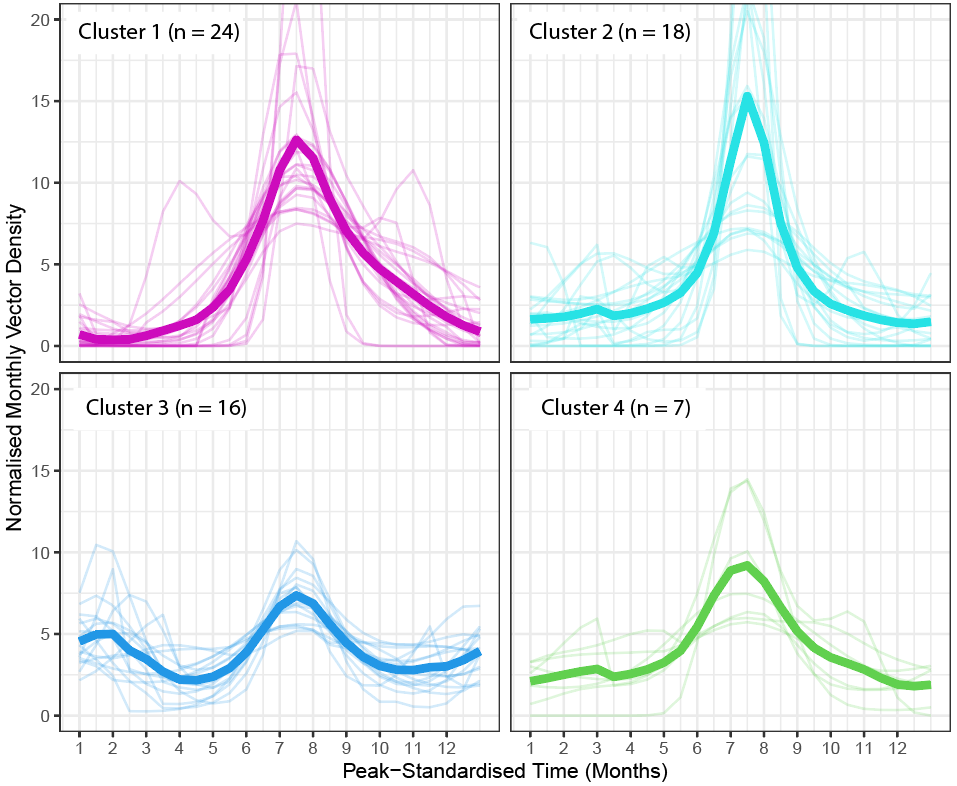


**Supplementary Information 3: Additional Figures and Results**

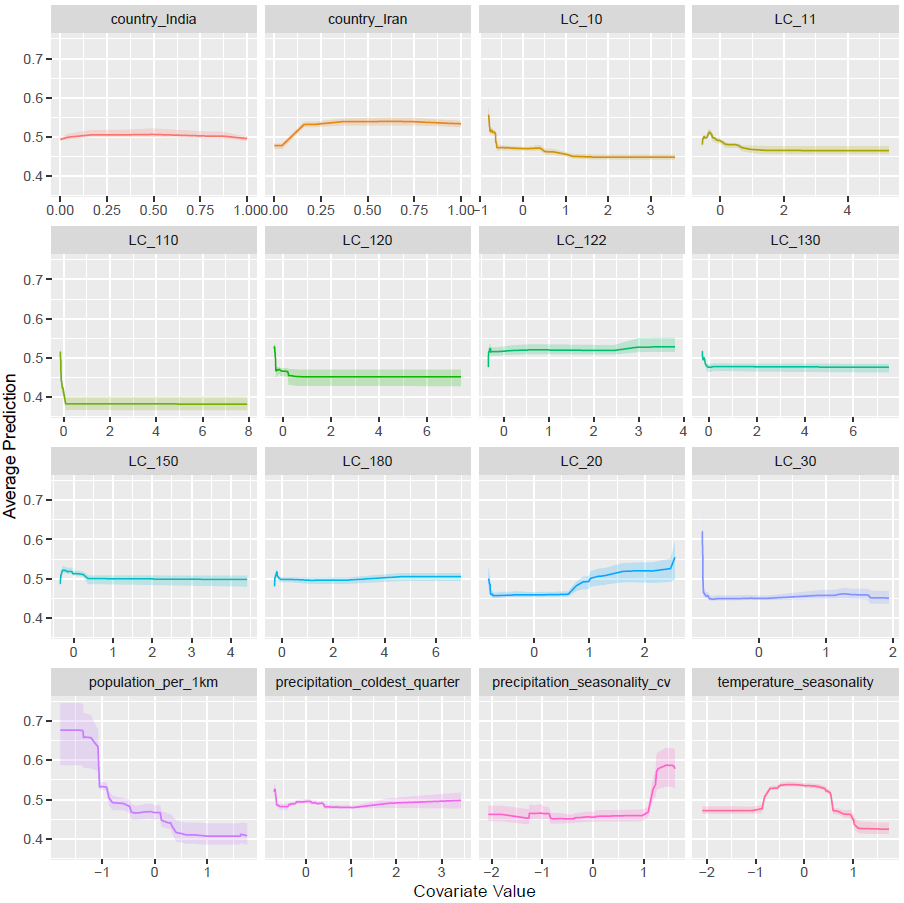
**Supplementary Figure 1: Results of model fitting to the longitudinal entomological data collated in this study.** Reviews of the literature in tandem with previously published databases of entomological data identified 65 *Anopheles stephensi* time-series matching the inclusion criteria (>10 months of catch data at monthly temporal resolution or finer), and a negative binomial gaussian process with period kernel fitted to each time-series. For the results presented above, black points are the data, and the lines represent the model output, coloured according to the country in which the study was conducted. Line indicates the mean model output, with the shaded ribbon delineating the 95% credible interval (CI).



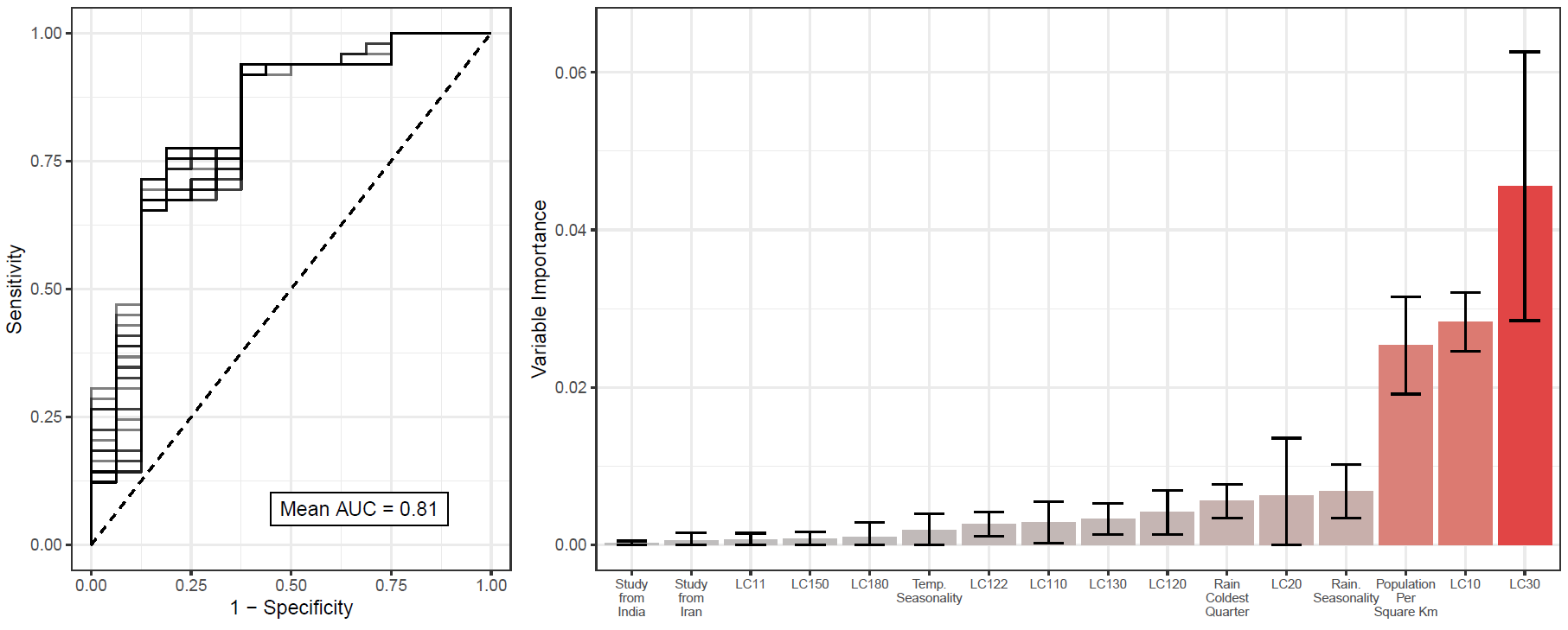
**Supplementary Figure 2: Archetype/Cluster Temporal Properties.** A series of mathematical operations were applied to the fitted time-series to characterise and explore their temporal properties. The results of this characterisation were then clustered using the k-means algorithm. For each cluster, the mean temporal profile is displayed, as well as the underlying distribution of values for each temporal property. For further information on each of these operations, see ***Supplementary Information: Time Series Characterisation and Analysis.***



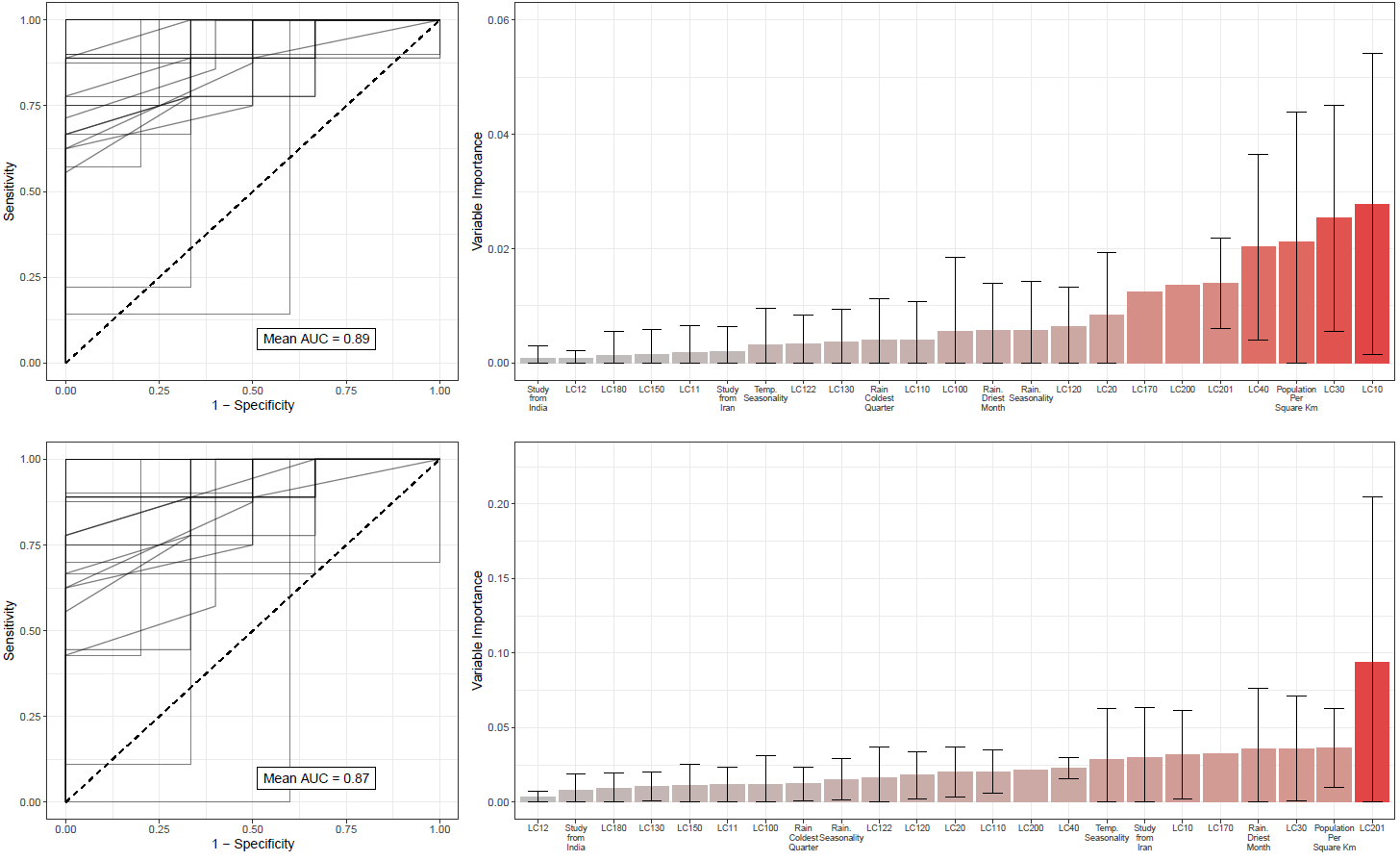
**Supplementary Figure 3: Results of Clustering For 4 Clusters Instead of 2.** In order to further investigate the different patterns of temporal dynamics present in the collated dataset, we re-ran the k-means clustering algorithm this time specifying 4 clusters. The less seasonal cluster from the 2 cluster analysis in the main text (Cluster 2 in the main text results) was retained (here Cluster 3), and Cluster 1 from the main text was further disaggregated into 3 different clusters (here, Clusters 1, 2 and 4), each defined by different peak timings (mean timing of vector density peak 7, 8.25 and 5.86 months after January for Clusters 1, 2 and 4 respectively) and the timing of the vector peak relative to peaks in rainfall (rainfall peak on average 1.03 and 2.32 months before vector density peak for Clusters 1 and 2, 1.09 months after vector density peak on average for Cluster 4).



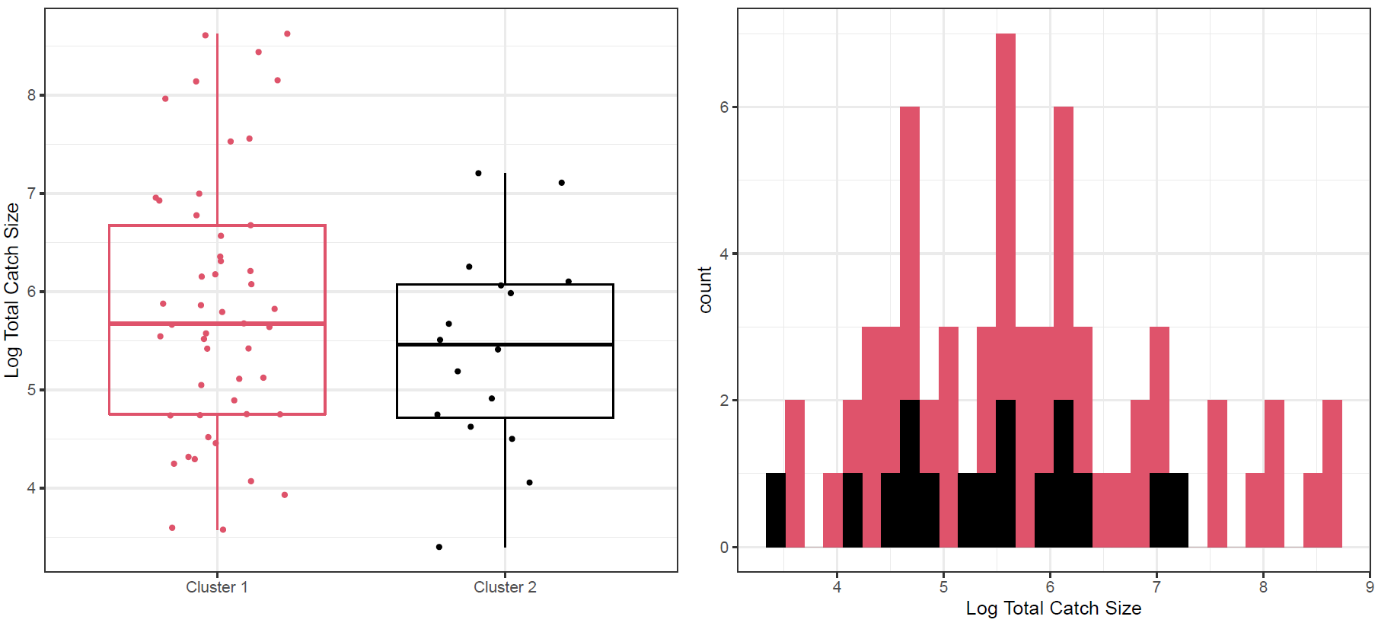
**Supplementary Figure 4: Partial Dependence Plots for Covariates Used in the Random Forest Classification Modelling.** The y-axis on the left shows the probability of the time-series belonging to Cluster 2 (i.e. a high probability indicates the time-series is predicted to likely belong to Cluster 2, a low probability indicates the time-series likely belongs to Cluster 1).The x-axis describes the value of the (scaled, normalised) covariate.



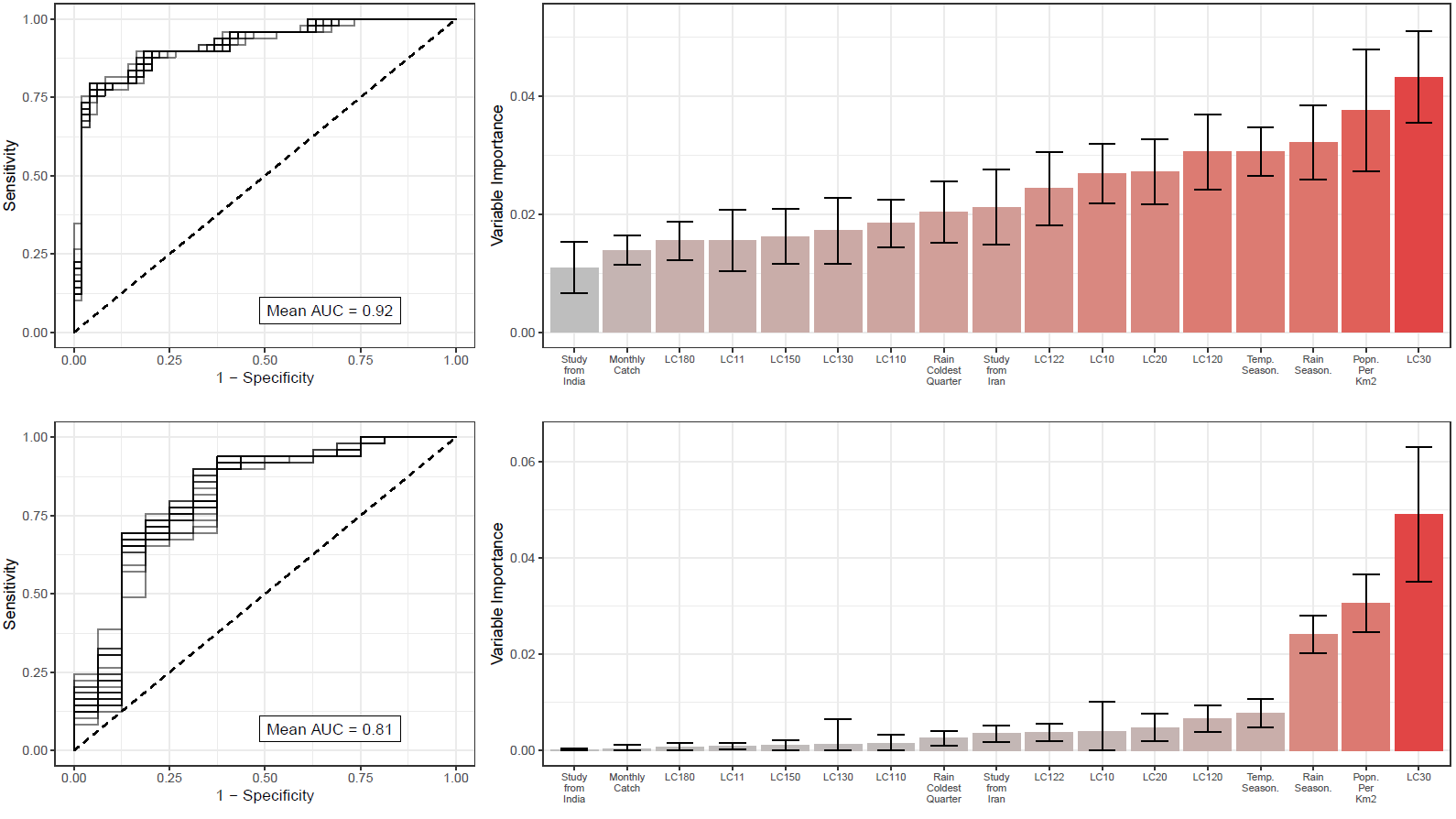
**Supplementary Figure 5: Random Forest Classification Results Without Upsampling Cluster 2.** Due to the extreme class-imbalance of Clusters 1 and 2 (49 vs 16 time-series respectively), the results presented in the main text are following upsampling of the Cluster 2 time-series to create a dataset with equal numbers of time-series belonging to each cluster. As a sensitivity analysis, we also carried out the random forest fitting without upsampling, and assessed both model fit (as measured by AUC) and variable importance. Whilst model fit was somewhat reduced compared to the upsampled data (mean AUC of 0.81 vs mean AUC >0.9 for the upsampled dataset), variable importance results were broadly consistent across both analyses, with population per square kilometre and various land-cover measures all emerging as important predictive variables.



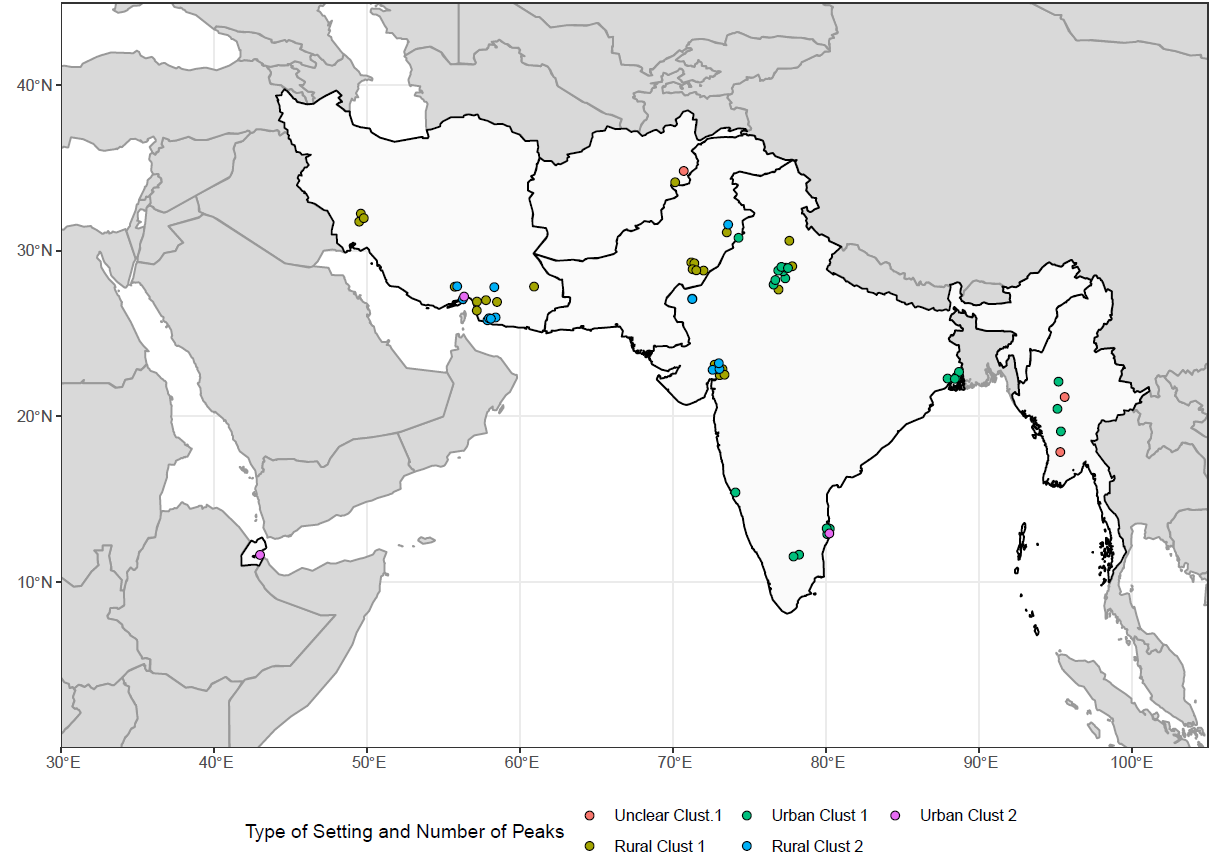
**Supplementary Figure 6: Random Forest Classification Results With Hold-Out Data.** Due to the overall sample size (n = 65 time-series), the results presented in the main text were generated using a random forest-based workflow where final model fitting (using hyperparameters tuned using 6-fold cross-validation) utilised the entirety of the dataset. As a sensitivity analysis, we also carried out the random forest fitting holding out a small portion of the dataset (n = 9) during model fitting, with model performance subsequently evaluated on this held-out data. Results presented above are in the case where data was upsampled to address class imbalance (top) and where no upsampling was carried out (bottom).



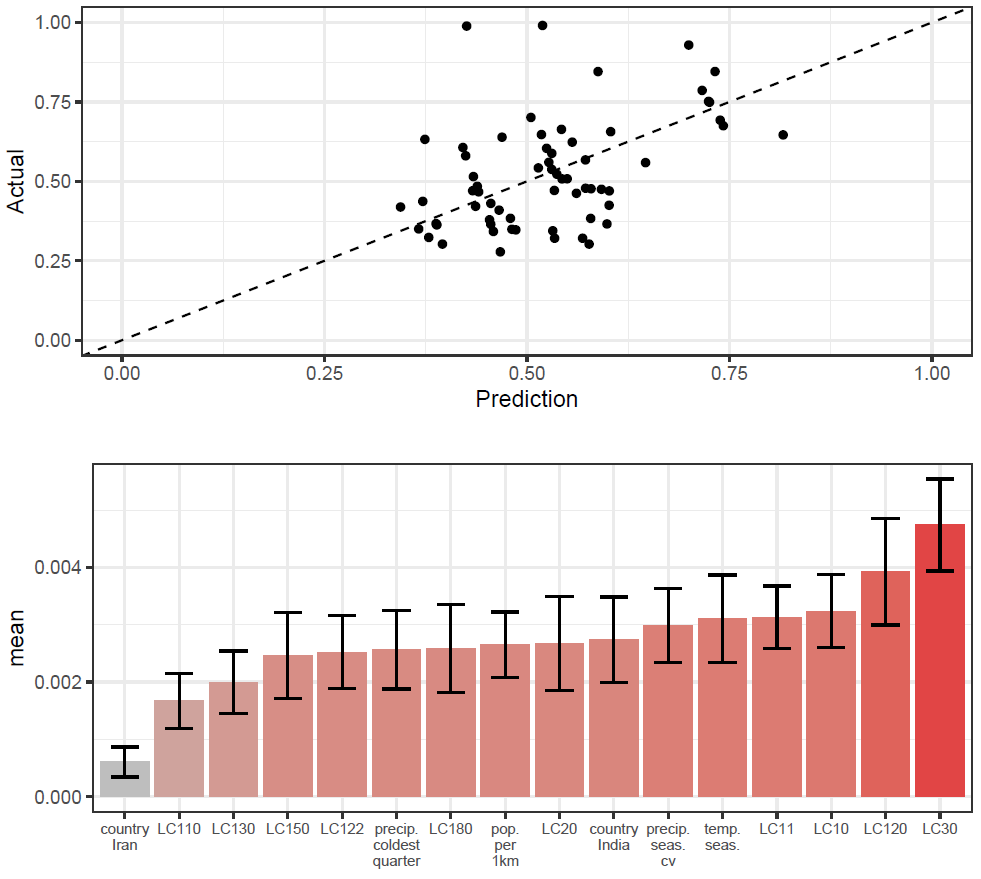
**Supplementary Figure 7: Exploring Variation In Total Catch Size By Cluster.** Boxplot and histogram of the total number of *Anopheles stephensi* mosquitoes caught over the duration of each study, coloured according to cluster membership. Total catch size was highly overdispersed, varying over several orders of magnitude across both clusters. There was no significant difference in the median catch size across clusters (Mood’s Median Test, p=0.4).



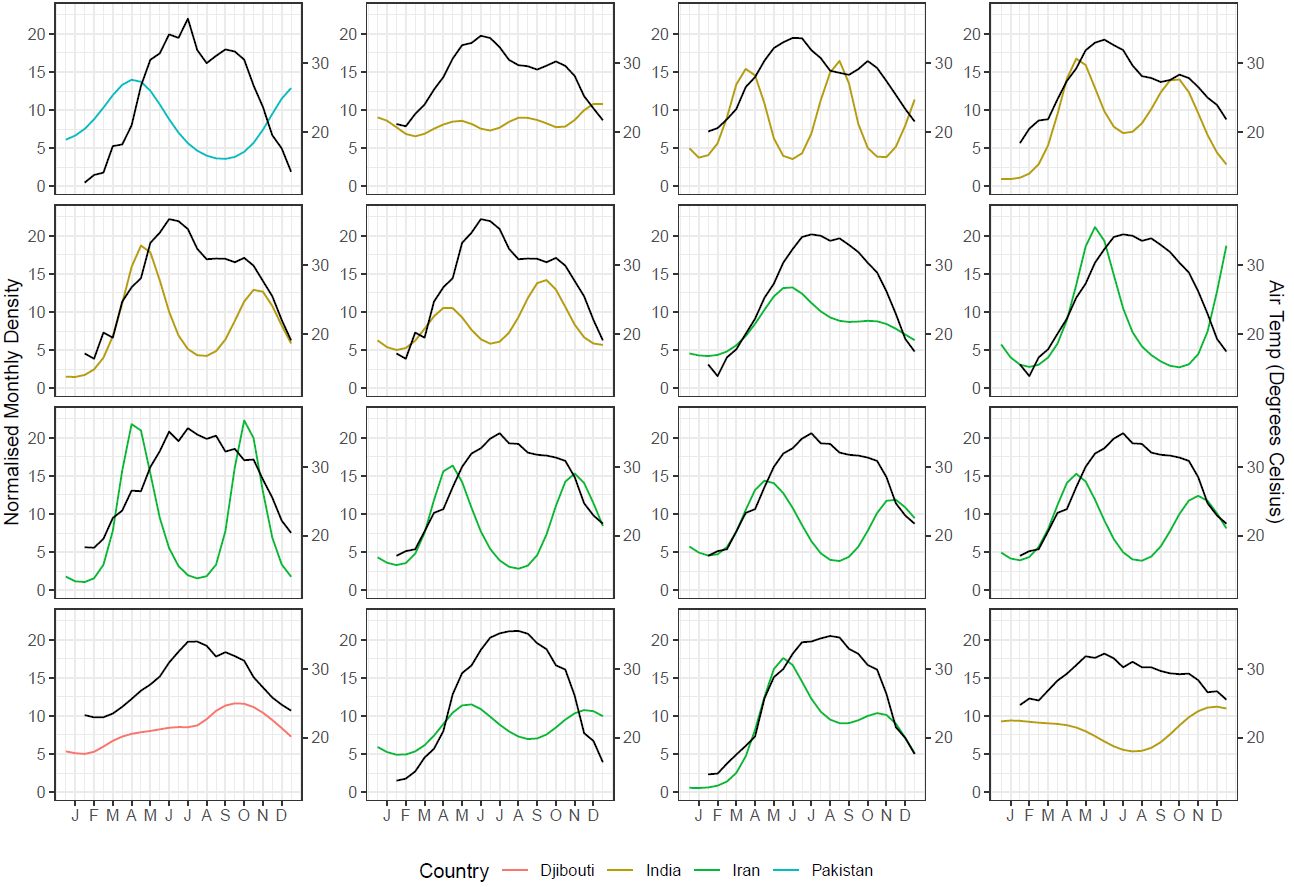
**Supplementary Figure 8: Random Forest Classification Results Including Monthly Catch Size As A Model Covariate.** As a sensitivity analysis, we also carried out the random forest fitting including average monthly catch size for each time-series as a predictive covariate, to assess whether any of the cluster assignments might be due to the model spuriously not detecting seasonal peaks in studies with low catch sizes. Results presented above are in the case where data was upsampled to address class imbalance (top) and where no upsampling was carried out (bottom), with predictive performance and variable importance largely changed compared to results presented in the main text (which do not include average monthly catch size as a covariate).



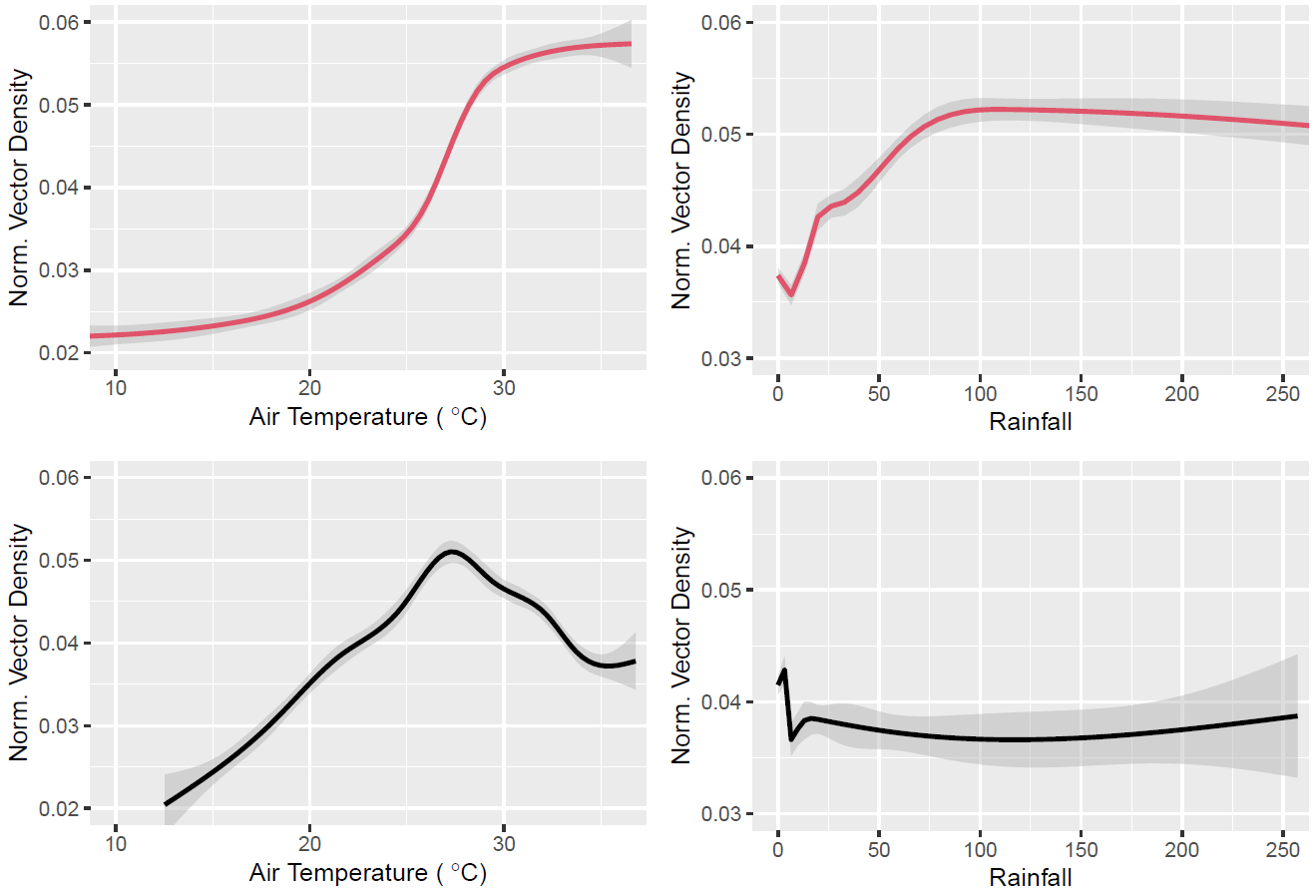
**Supplementary Figure 9: Sources and Locations of *Anopheles stephensi* Time-Series Data According to Urban/Rural Assignment.** Collated time-series are displayed above coloured according to 1) whether or not the study was carried out in an urban or rural location; and 2) which cluster they were assigned to.



**Supplementary Figure 10: Random Forest Prediction of Percentage of Vector Density In Any 3 Month Period.** As a further sensitivity analysis, we used a random forest modelling approach to predict the percentage of vector density occurring in a single continuous 3-month period. Results presented above are the average of 25 independents random forest model fittings, with no upsampling of the data carried out, and the final model fitted (using hyperparameters tuned using 6-fold cross-validation) to the full dataset. Model predictive power was limited, with correlation between predicted and actual values = 0.43.



**Supplementary Figure 11: Exploratory Analysis of 2m Air Temperature and Vector Density for Time-Series Belonging to Cluster 2.** In all plots, coloured line indicates mean vector density from the fitted negative binomial gaussian process model, coloured by country, and the black line indicates average 2m air temperature over the area and for the year the study was carried out in, lagged by 1 month relative to vector density.



**Supplementary Figure 12: Calculated Cluster Specific Relationship Between Rainfall/Air Temperature and Vector Density By Cluster.** For all time-series belonging to each cluster, the normalised vector density was plotted against the rainfall and 2m air temperature, and a smooth loess curve fitted. Top panels display the univariate relationships between the two covariates and normalised vector density for Cluster 1 time-series, bottom panels display the univariate relationships for Cluster 2 time-series.

**References**

1. Sinka, M. E. *et al.* A new malaria vector in Africa: Predicting the expansion range of Anopheles stephensi and identifying the urban populations at risk. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 24900–24908 (2020).

2. Whittaker, C. *et al.* The ecological structure of mosquito population seasonal dynamics. *bioRxiv* (2021) doi:10.1101/2021.01.09.21249456.

3. Fulcher, B. D., Little, M. A. & Jones, N. S. Highly comparative time-series analysis: the empirical structure of time series and their methods. *J. R. Soc. Interface* **10**, 20130048 (2013).

4. Fick, S. E. & Hijmans, R. J. WorldClim 2: new 1‐km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302–4315 (2017).

5. European Space Agency (ESA). ESA CCI Land Cover time-series v2.0.7 (1992 - 2015).

6. Gaughan, A. E., Stevens, F. R., Linard, C., Jia, P. & Tatem, A. J. High resolution population distribution maps for Southeast Asia in 2010 and 2015. *PLoS One* **8**, e55882 (2013).

7. Linard, C., Gilbert, M., Snow, R. W., Noor, A. M. & Tatem, A. J. Population distribution, settlement patterns and accessibility across Africa in 2010. *PLoS One* **7**, e31743 (2012).

8. Justice, C. O. *et al.* An overview of MODIS Land data processing and product status. *Remote Sens. Environ.* **83**, 3–15 (2002).

9. Didan, K. MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006 NASA EOSDIS Land Processes DAAC. doi:10.5067/MODIS/MOD13Q1.006.

10. Kuhn, M. & Wickham, H. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. *Boston, MA, USA. [(accessed on 10 December 2020)]* (2020).

11. Funk, C. *et al.* The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data* **2**, 1–21 (2015).

12. Rasmussen, C. E. Gaussian processes in machine learning. in *Advanced Lectures on Machine Learning* 63–71 (Springer Berlin Heidelberg, 2004).

13. Carpenter, B. *et al.* Stan: A probabilistic programming language. *J. Stat. Softw.* **76**, 1–32 (2017).

14. Cairns, M. *et al.* Estimating the potential public health impact of seasonal malaria chemoprevention in African children. *Nat. Commun.* **3**, 881 (2012).

15. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

16. Biau, G. Analysis of a random forests model. *J. Mach. Learn. Res.* **13**, 1063–1095 (2012).

17. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *arXiv [stat.ML]* (2015).

18. Molnar, C. *Interpretable Machine Learning*. (Lulu.com, 2020).

19. Griffin, J. T., Ferguson, N. M. & Ghani, A. C. Estimates of the changing age-burden of Plasmodium falciparum malaria disease in sub-Saharan Africa. *Nat. Commun.* **5**, 3136 (2014).

20. Challenger, J. D. *et al.* Predicting the public health impact of a malaria transmission-blocking vaccine. *Nat. Commun.* **12**, 1494 (2021).

21. Griffin, J. T. *et al.* Reducing Plasmodium falciparum malaria transmission in Africa: a model-based evaluation of intervention strategies. *PLoS Med.* **7**, e1000324 (2010).

22. White, M. T. *et al.* Modelling the impact of vector control interventions on Anopheles gambiae population dynamics. *Parasit. Vectors* **4**, 153 (2011).

23. Hamlet, A. *et al.* The potential impact of Anopheles stephensi establishment on the transmission of Plasmodium falciparum in Ethiopia and prospective control measures. *bioRxiv* (2021) doi:10.1101/2021.08.19.21262272.