The Ecological Structure of Mosquito Population Dynamics: Insights from India, Consequences for Malaria Control

Charlie Whittaker1\*, Aaron Morris1, Peter Winskill1, Benjamin Lambert1, Marianne Sinka2, Daniel Weiss3, Ashwani Kumar4, Azra Ghani1\* & Samir Bhatt1,3\*

1Department of Infectious Disease Epidemiology, Imperial College, London, UK

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

3Malaria Atlas Project, Oxford Big Data Institute, Nuffield Department of Medicine, University of Oxford, Oxford, UK

4Insert Ashwani’s affiliation here

\*Corresponding Author: [charles.whittaker16@imperial.ac.uk](mailto:charles.whittaker16@imperial.ac.uk)

\*Denotes joint-authorship

**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**

**Supplementary Information 4: List of 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* mosquito population dynamics across the Indian sub-continent. In Supplementary Information 1, we present an overview of the systematic search strategy employed, as well as details about the data collated and the initial pre-processing applied to it. In Supplementary Information 2, we detail the statistical methodologies employed to process this extracted data, methodologies whose output forms the basis for the results presented in the main text. Finally, in Supplementary Information 3, we present an array of figures 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**

Web of Science and PubMed databases were searched on 17th October 2017 using the keywords “India” AND “Anophel\*” in order to identify references containing temporally disaggregated entomological data. Our searches identified a total of 1945 records, with 1556 remaining after duplicate removal. References were selected for Inclusion/Exclusion according to the following criteria:

**Inclusion Criteria:**

* Reference contains temporally disaggregated adult mosquito catch data at a temporal resolution of at least monthly.

**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).
* Reference only contains information on immature/larval mosquito life cycle stages.
* Reference contained insufficient information to geolocate the area in which the study was conducted.

Following Title and Abstract screening, a total of 281 records were identified, with 78 references retained after Full Text Evaluation. These 78 references contained temporally disaggregated *Anopheles* catch information at monthly resolution (no references were identified which presented higher temporal resolution catch data) for a total of 106 distinct and geolocatable areas across the Indian subcontinent. These form the basis for the results presented in this paper. The next section goes into 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 detailed. We restricted extraction to 7 major *Anopheles* species known to be relevant to malaria transmission in India (although a number of others exist) and for which multiple catch data time series were available. These were *Anopheles annularis*1, *Anopheles culicifacies*2,3, *Anopheles dirus*4,5, *Anopheles fluviatilis*6,7, *Anopheles minimus*8,9, *Anopheles stephensi*10 and *Anopheles subpictus*3,11. Where data were presented in the form of a table, data was copied directly from the table. Where graphs only were presented, estimates of the data were extracted using DataThiefTM software. This yielded a total of 305 time series of monthly mosquito catch data, ranging in length from 5 – 46 months. We restricted subsequent analyses to time series that spanned a year (12 timepoints, monthly) or longer, a total of 294 time series. This yielded the following number of time series for each of the species considered (Supplementary Table 1):

**Supplementary Table 1: Summary of the Number of Time Series Extracted, Disaggregated By Species**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | ***A.annularis*** | ***A.culicifacies*** | ***A.dirus*** | ***A.fluviatilis*** | ***A.minimus*** | ***A.stephensi*** | ***A.subpictus*** |
| **# Time Series** | 44 | 96 | 15 | 66 | 14 | 28 | 42 |
| **>= 12 Months** | 42 | 94 | 14 | 62 | 13 | 28 | 41 |

As the primary focus of this research was to explore annual and seasonal patterns of mosquito population dynamics, as well as the fact that variations in time series length are a factor known to affect their statistical properties (and which would therefore impact the comparability of the time series gathered and analysed here) (ref from Oxford Time Series Supplementary Information), 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. 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 studies analysed here employed a wide array of different sampling methodologies including Indoor and Outdoor Resting Collections, Human Landing Catches, Spray Catches and Trap Catches amongst others. Results were typically, though not always, presented in the form of some sampling-effort standardised measure such as Man Hour Density (MHD). As such, though reflective of mosquito population dynamics, these measures do not represent the overall number of mosquitoes caught. To this end, where information on sampling effort (number of hours spent sampling, number of households/cattlesheds searched, number of human baits, number of traps set etc) was present, we used this information to convert MHD back to the raw counts. In the small number instances where there was variable sampling effort across the time series (which would bias the conversion away from the true underlying population dynamics), we conservatively used the lowest sampling effort recorded across the time series. 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: Temporal Information, sheet Raw Catch Data for more information about the transformations applied to each of the time series, as well as the unprocessed catch data extracted from each of the references.

**Environmental Covariate Assembly**

The environmental covariates (i.e. the independent variables used to predict the different seasonal patterns) used in this research consist of raster layers spanning all of India at a 2.5 arc-minute (~ 5km by 5km) spatial resolution. The covariates utilised here were initially selected based on those previously used in other mosquito mapping efforts (#s 1 – 30 in Supplementary Table 2), which were focussed on evaluating species Presence/Absence rather than temporal dynamics. Additionally, after consideration of some of the possible drivers of seasonal dynamics (primarily hydrological considerations surrounding the seasonality and availability of aquatic breeding sources, and how this might interact with environmental composition12,13 and species specific breeding preferences14,15 to structure population dynamics), a number of other raster layers were included that together describe further the underlying hydrological environment (#s 31 – 40 in Supplementary Table 2). The majority of these covariates are derived from high temporal resolution satellite images that were initially gap-filled16 to eliminate missing data that typically arises from cloud cover. These images were then aggregated and summarised to produce the suite of synoptic environmental covariates utilised here. **NOTE NEED TO HAVE A THINK ABOUT MONTHLY VS SYNOPTIC, AND ADDITIONALLY, WHETHER THERE ARE ANY MORE COVARIATES I NEED TO THINK ABOUT INCLUDING.**

**Study Geolocation**

Geolocation of study areas was possible to a varying degree depending on the information available within the paper (and related literature). When villages names or the details of the administrative unit a study was carried out in were provided in the paper text, geolocation was carried out utilising a wide array of resources containing spatially explicit information on the location of Indian settlements and administrative units. These were Google Maps/Google Earth, Etrace, OneFiveNine, Veethi, Wikimapia, VillageInfo, MapsOfIndia, Geonames and AlipurduarTourism. Additionally, a number of the references identified in our review had previously been utilised as part of the Malaria Atlas Project (MAP) Presence/Absence mapping work and so had previously been geolocated17. In these instances, the MAP location estimate was used. The precision of study location estimates varied greatly (due to the extent of spatial detailed provided in the paper e.g. village vs district as well as the identifiability of villages/administrative units) – this uncertainty is explicitly incorporated into our analyses, with raster covariates extracted over the full area the study is believed to have been carried out in. See Supplementary Data Overall Temporal Information, sheet Location & Spatial Information for more information about the specific resources used to geolocate each individual study location.

**Rainfall Data Extraction**

For each of the 108 study locations identified and geolocated, daily rainfall estimates spanning the sampling period were also collated. These data were taken from “The Climate Hazards Group Infrared Precipitation With Stations” (CHIRPS) dataset18 and were subsequently aggregated up to the monthly level in order to facilitate comparison and correlation assessment with the monthly mosquito catch data.

**Supplementary Table 2: Environmental Covariates Explored in the Variable Selection Process and Utilised in Modelling and Prediction of Seasonal Population Dynamics**

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Variable** | **Temporal Resolution** | **Source** |
| 1 | Annual Mean Temperature | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 2 | Mean Diurnal Range | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 3 | Isothermality | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 4 | Temperature Seasonality | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 5 | Max Temperature of Warmest Month | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 6 | Min Temperature of Coldest Month | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 7 | Temperature Annual Range | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 8 | Mean Temperature of Wettest Quarter | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 9 | Mean Temperature of Driest Quarter | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 10 | Mean Temperature of Warmest Quarter | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 11 | Mean Temperature of Coldest Quarter | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 12 | Annual Precipitation | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 13 | Precipitation of Wettest Month | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 14 | Precipitation of Driest Month | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 15 | Precipitation Seasonality | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 16 | Precipitation of Wettest Quarter | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 17 | Precipitation of Driest Quarter | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 18 | Precipitation of Warmest Quarter | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 19 | Precipitation of Coldest Quarter | Annual Average, 1970 - 2000 | <https://www.worldclim.org/bioclim> |
| 20 | Potential Evapotranspiration | Annual Average, 1950 - 2000 | <https://cgiarcsi.community/data/global-aridity-and-pet-database/> |
| 21 | Global Aridity Index | Annual Average, 1950 - 2000 | <https://cgiarcsi.community/data/global-aridity-and-pet-database/> |
| 22 | Population Density | 2010 | <http://www.worldpop.org.uk> |
| 23 | Day Land Surface Temperature Mean | Annual Average 2001 – 2015 | <https://developers.google.com/earth-engine/datasets/catalog/Oxford_MAP_LST_Day_5km_Monthly> |
| 24 | Day Land Surface Temperature SD | Annual Average 2001 – 2015 | <https://developers.google.com/earth-engine/datasets/catalog/Oxford_MAP_LST_Day_5km_Monthly> |
| 25 | Night Land Surface Temperature Mean | Annual Average 2001 – 2015 | <https://developers.google.com/earth-engine/datasets/catalog/Oxford_MAP_LST_Night_5km_Monthly> |
| 26 | Night Land Surface Temperature SD | Annual Average 2001 – 2015 | <https://developers.google.com/earth-engine/datasets/catalog/Oxford_MAP_LST_Day_5km_Monthly> |
| 27 | Tasselled Cap Wetness Mean | Annual Average 2001 – 2012 | <https://developers.google.com/earth-engine/datasets/catalog/Oxford_MAP_TCW_5km_Monthly> |
| 28 | Tasselled Cap Wetness SD | Annual Average 2001 – 2012 | <https://developers.google.com/earth-engine/datasets/catalog/Oxford_MAP_TCW_5km_Monthly> |
| 29 | Tasselled Cap Brightness Mean | Annual Average 2001 – 2012 | <https://developers.google.com/earth-engine/datasets/catalog/Oxford_MAP_TCB_5km_Monthly> |
| 30 | Tasselled Cap Brightness SD | Annual Average 2001 – 2012 | <https://developers.google.com/earth-engine/datasets/catalog/Oxford_MAP_TCB_5km_Monthly> |
| 31 | Elevation | NA | <https://developers.google.com/earth-engine/datasets/catalog/USGS_SRTMGL1_003> |
| 32 | Flow Accumulation | NA | <https://developers.google.com/earth-engine/datasets/catalog/WWF_HydroSHEDS_15ACC> |
| 33 | Specific Humidity Mean | Annual Average 1948 - 2010 | <https://developers.google.com/earth-engine/datasets/catalog/NASA_GLDAS_V20_NOAH_G025_T3H> |
| 34 | Specific Humidity SD | Annual Average 1948 – 2010 | <https://developers.google.com/earth-engine/datasets/catalog/NASA_GLDAS_V20_NOAH_G025_T3H> |
| 35 | Enhanced Vegetation Index | Yearly Average 2001 – 2015 | <https://developers.google.com/earth-engine/datasets/catalog/Oxford_MAP_EVI_5km_Monthly> |
| 36 | Landcover | Yearly Average 2001 – 2012 | <https://developers.google.com/earth-engine/datasets/catalog/Oxford_MAP_IGBP_Fractional_Landcover_5km_Annual> |
| 37 | Water Body Maximum Extent | 1984 – 2015 | <https://developers.google.com/earth-engine/datasets/catalog/JRC_GSW1_0_GlobalSurfaceWater> |
| 38 | Water Body Seasonality | Average 1984 – 2015 | <https://developers.google.com/earth-engine/datasets/catalog/JRC_GSW1_0_GlobalSurfaceWater> |
| 39 | Water Areas Occurrence | Average 1984 – 2015 | <https://developers.google.com/earth-engine/datasets/catalog/JRC_GSW1_0_GlobalSurfaceWater> |
| 40 | Distance to Nearest Water Body | Average 1984 – 2015 | Generated manually using the Water Areas Occurrence raster |

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

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

We use a highly flexible class of stochastic models known as Gaussian Processes in order to temporally interpolate between the monthly catch datapoints and produce smoothed estimates of mosquito abundance spanning the entire year. Gaussian processes represent a distribution over functions such that any finite set of function values have a joint Gaussian distribution (Rasmussen reference). The Gaussian process is entirely specified by its mean function, defined as:

and by its covariance function:

also known as the kernel. This kernel is a positive-definite function of two inputs, and that defines the covariance between any two points (and by extension the covariance matrix of our Gaussian Process when all pairwise combinations of points are considered). In doing so, the kernel encodes prior information about the extent to which we would expect two objects ( and in this instance) to be similar. A wide array of kernels exist that specify an equally wide array of similarity structures, such as the squared exponential (where similarity varies with the Euclidean distance separating and ) and the linear kernel (which allows the relationship governing similarity to vary with not just the relative position of two inputs, i.e. , but with their absolute position, a property that makes this kernel “non-stationary”). Given the strong seasonality known to be present in mosquito catch time series (**Ref)** and from the empirically observed patterns of abundance observed when examining the raw time series **(Supplementary Figure 1)**, we selected a Periodic Kernel. This kernel defines similarity based on the distance between and compared to some period and so is able to accommodate patterns that broadly repeat themselves over time (such as seasonal or annual peaks in mosquito abundance).

Where the represents the period, specifies the magnitude of the covariance given a certain period, and represents a lengthscale parameter further constraining the extent to which two values separated by a given distance can co-vary with one another.

Bayesian inference and fitting of Gaussian Processes typically utilises the following hierarchical formulation:

where represents a vector of hyperparameters involved in defining the kernel’s properties, is a vector of values representing a realisation from Gaussian Process, and our observed counts. However, mosquito catch data is rarely normally distributed, frequently displaying high levels of overdispersion **(Ref)**, a common property of biological systems generally, but made more acute by the fact that for a number of the time series, the monthly catches reported represented the summed total of multiple catches made throughout the catch. This process of summation also introduces overdispersion **(Ref or something?)**. Motivated by this, we adapted the above framework to accommodate a Negative Binomial likelihood, leading to the following inferential framework:

where is the exponential function (), reflecting the fact that we use a log link between the observed counts and the underlying latent process reflecting the population dynamics.

**Prior Probability Specification**

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 period prior was centred around (a value which would represent annual variation) to reflect the fact that the majority of observed variation in mosquito abundance recorded has typically been observed to cycle annually **(REF, but also see Supplementary Figure 1)**. A wide standard deviation was used however in to allow the model to identify and accommodate instances of bimodality or periods operating across timescales longer than a year, although important to note is that the lower and upper bounds for the period were set to 4 and 18 months respectively, to avoid identifiability issues arising from the lack of data at temporal resolutions substantially below and above these bounds. A similarly wide prior was set over the overdispersion parameter and the scaling factor . An informative and tight prior was set for the lengthscale in order to constrain…..**HOW DO WE JUSTIFY THIS.**

Changes to the prior distribution did not significantly change posterior fits. Specifically, specifications using a less informative prior for the lengthscale did not significantly alter the qualitative conclusions drawn here, highlighting the robustness of the presented results.

**Model Fitting and Parameter Inference**

This Negative Binomial Gaussian Process were fitted using STAN, a probabilistic programming language for statistical inference written in C++ that employs gradient-based Markov Chain Monte Carlo algorithms (typically the No-U-Turn sampler, a variant of Hamiltionian Monte Carlo) for Bayesian inference **(Ref)**. The model specified above was implemented in R using the rStan package **(Ref)**. For each time series, 4 chains of 5,000 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 10,000 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, indicating stability of the chains and the probably convergence to the underlying true posterior distribution.

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

Following this fitting process, and to establish comparability across the time series (which varied substantially in the absolute count numbers recorded, we normalised each time series in the following way:

where is the normalised count for timepoint and is the un-normalised count taken from the Negative Binomial GP fitting described in the previous section.

To further characterise the periodic properties of these time series, we fit a Von Mises distribution, which is a continuous probability distribution on the circle with range from to . Broadly, it can be regarded as the circular analogue of the normal distribution 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 fit two sets of Von Mises densities to the normalised time series, the first containing a single component, specified as:

And the other possessing two components, formulated as:

where in both instances represents the normalised monthly count formulated as a random variable on the circle, i.e. by defining . Fitting was carried out in R using the *optim* function and with the sum of squares as the loss function. The outputs arising from this fitting – the comparative suitability of the one and two component distributions, as well as the values of , and , were then explored to further characterise the temporal properties of the data.

**Time Series Characterisation and Analysis:**

Following this, we sought to characterise the properties and features of the smoothed time series. To do this, we apply a series of mathematical operations to the time series to characterise and inform on features of interest. In doing so, we generate reduced dimensional representation of our time series that facilitate comparison of statistical properties and identification of time series that share similar statistical properties. The mathematical operations applied to the Negative Binomial GP fitted, normalised time series were the following:

1. **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. **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 inference based on the 10,000 MCMC iterations described in further detail in the section **Model Fitting and Parameter Inference.** The median value of across these 10,000 iterations was used here.
2. **Proportion of Points Greater Than 2x the Mean:** For each fitted, normalised time series, the proportion of points greater than 2x the mean of the time series was calculated. This informs about the extent to which the data is peaked, as well as the width of the peak.
3. **Peak Distance from January:** For each fitted, normalised time series, the maximum recorded value was noted and the distance of this value from January was calculated.
4. **Von Mises 1 Component Mean:** Following fitting of both 1 and 2 component Von Mises distributions to the Negative Binomial GP fitted, normalised time series (detailed further in the section **Von Mises Distribution Fitting**). If the 1 component Von Mises distribution was preferred, then the maximum likelihood of the mean is 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. **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, a time series was deemed to possess one peak if the value of the 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.
6. **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.

**Principal Components Analysis and Clustering:**

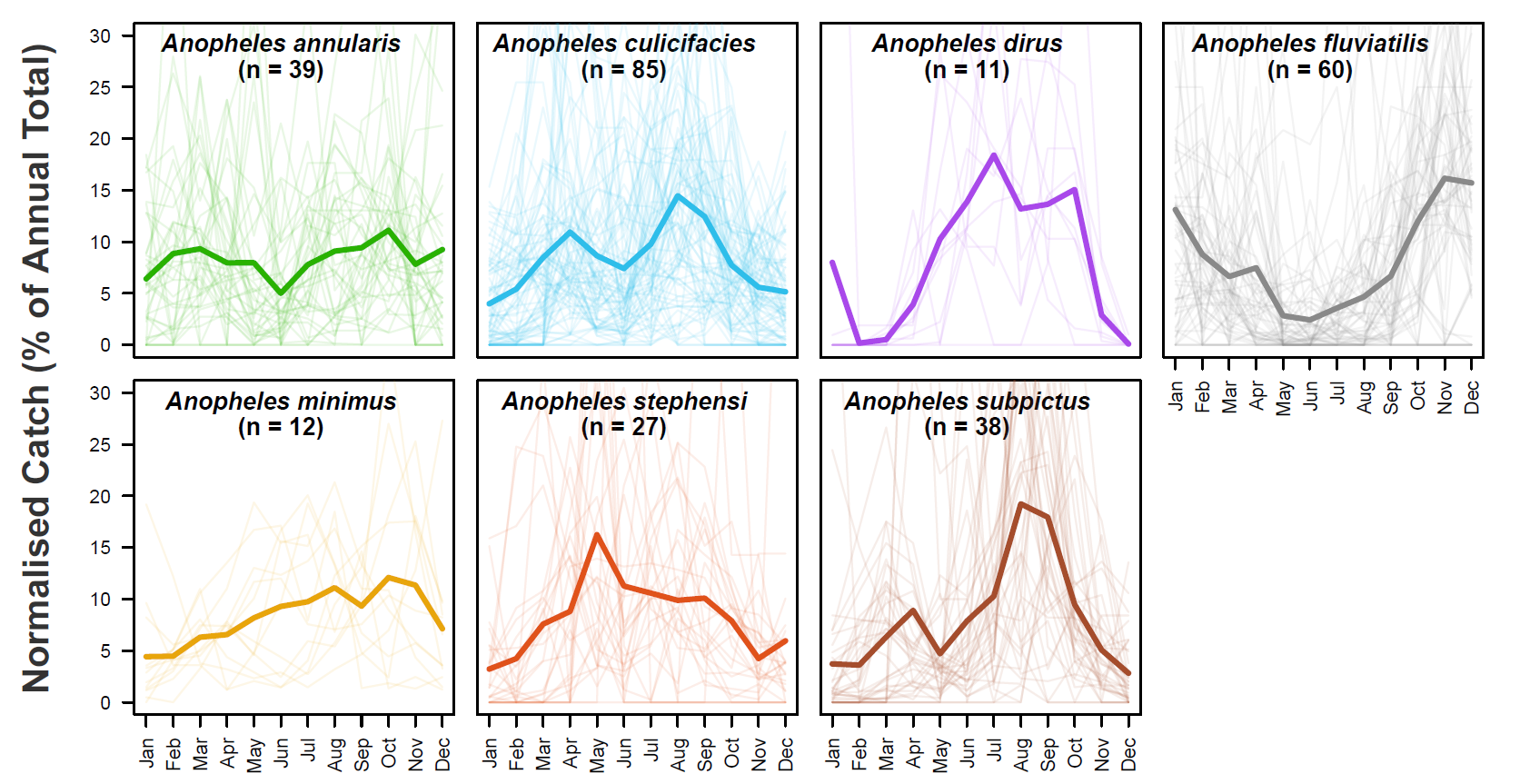
PCA is a statistical procedure that utilises an orthogonal transformation to convert a set of correlated variables (in this case the outputs of the 7 mathematical operations described above for each of the time series) into a set of linearly uncorrelated variables (known as the “principal components”). 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). To cluster the time series, we used the k-means clustering algorithm, which partitions the observations into k clusters, with each observation assigned to the cluster with the nearest mean.

**Penalised Logistic Regression Modelling:**

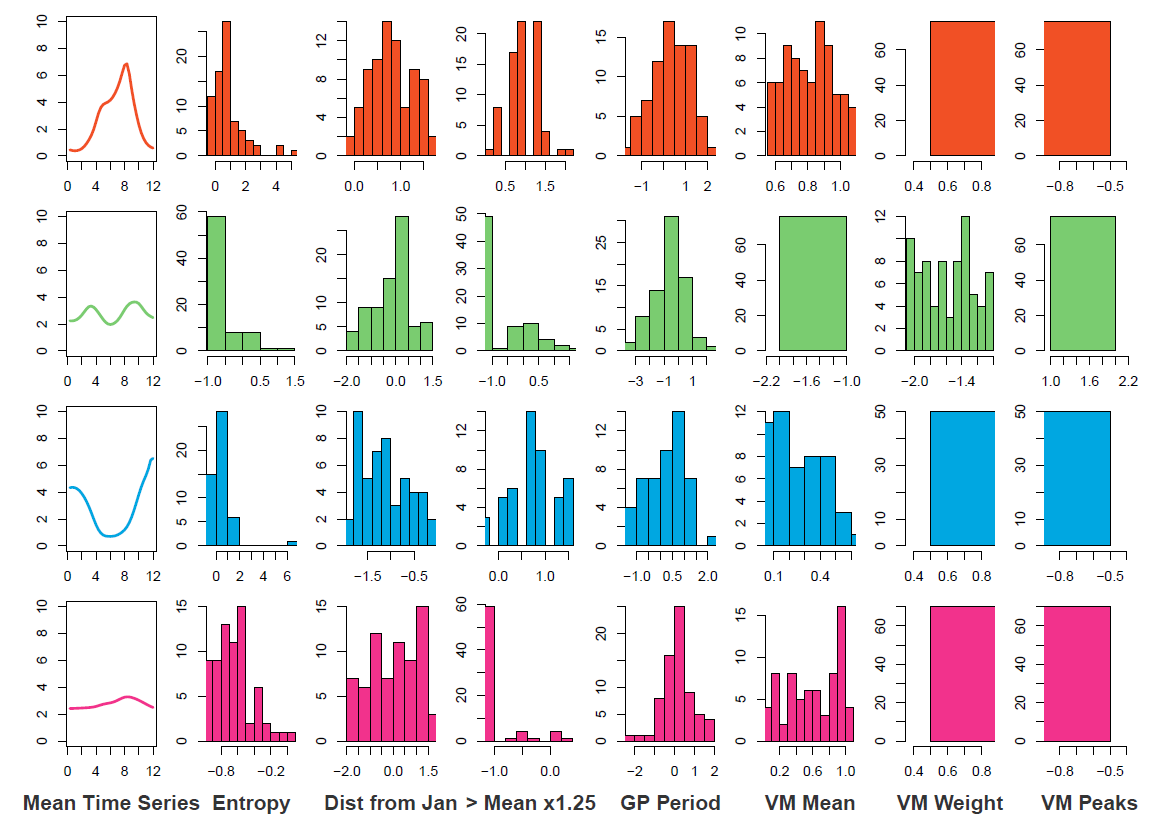
(get the bulk of this from Introduction to Statistical Learning)

**Evaluation of Model Accuracy:**

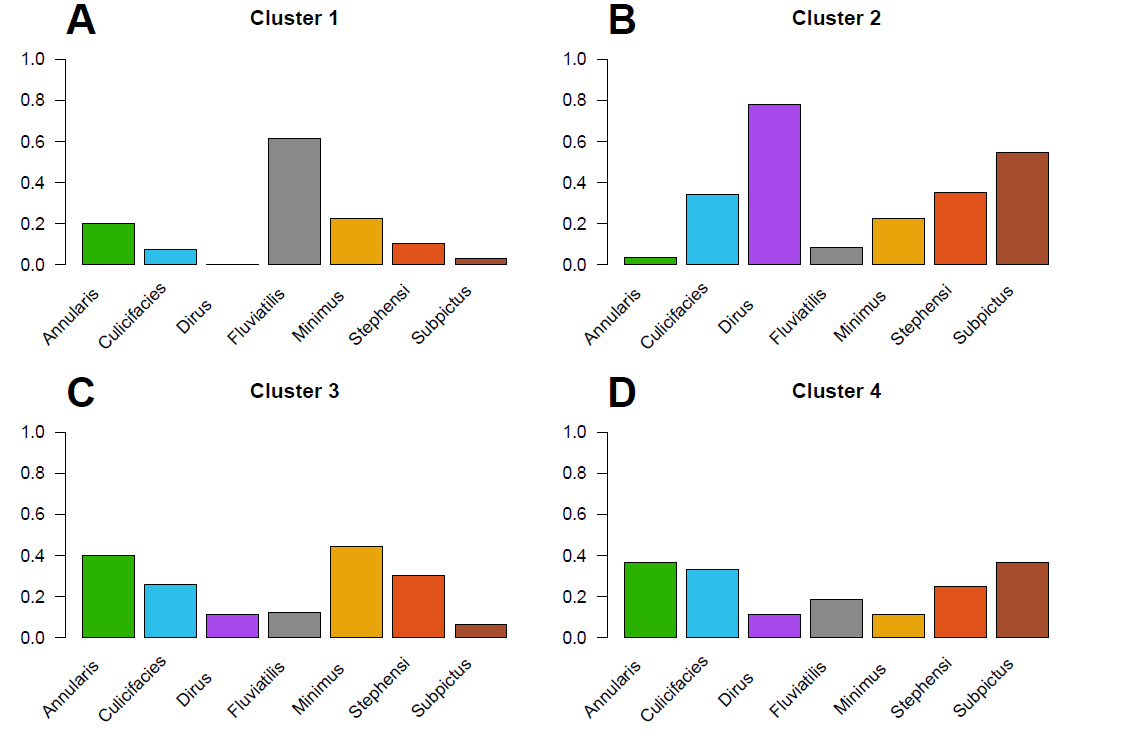
**Supplementary Information 3: Additional Figures and Results**



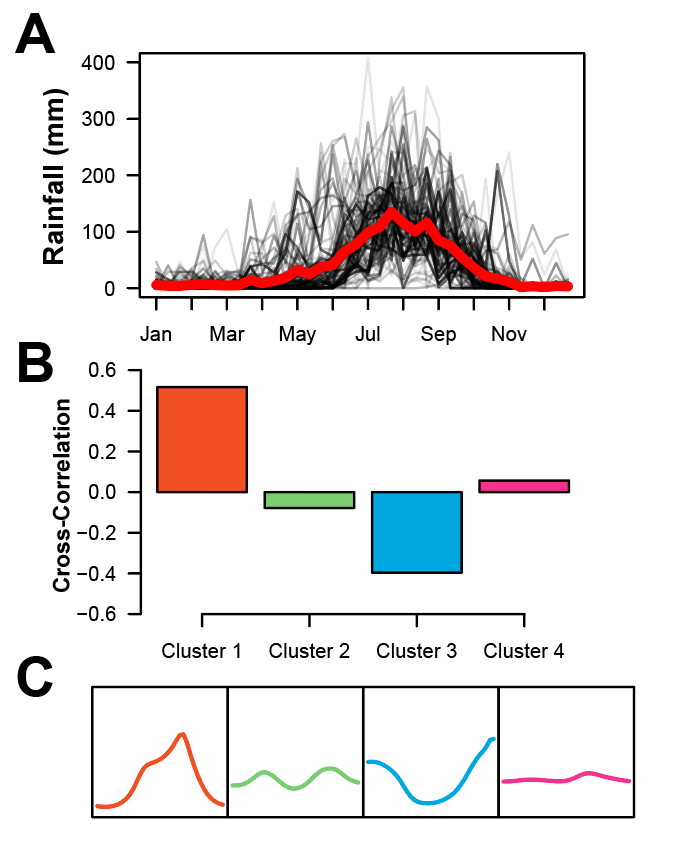
**Supplementary Figure 1: Raw Data Species Seasonal Plots**



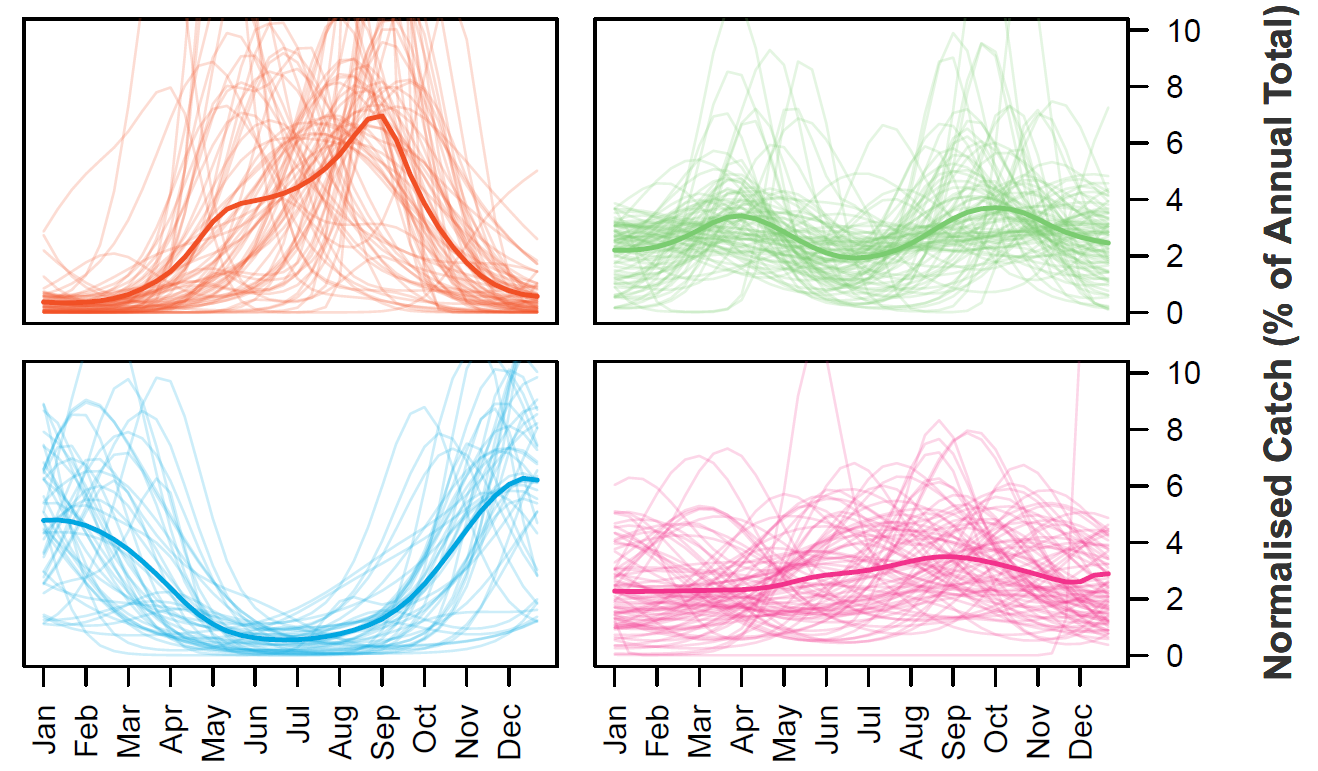
**Supplementary Figure 2: Cluster Temporal Properties**



**Supplementary Figure 3: Cluster Species Proportions**



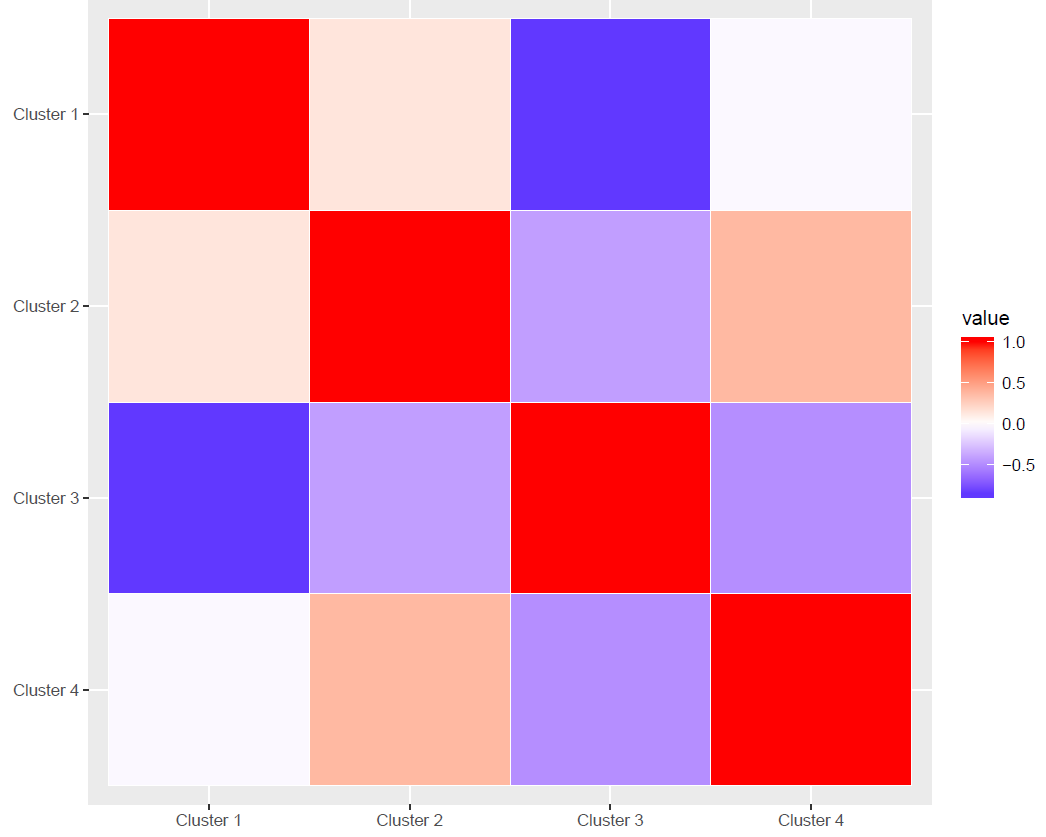
**Supplementary Figure 4: Rainfall Correlation Results**



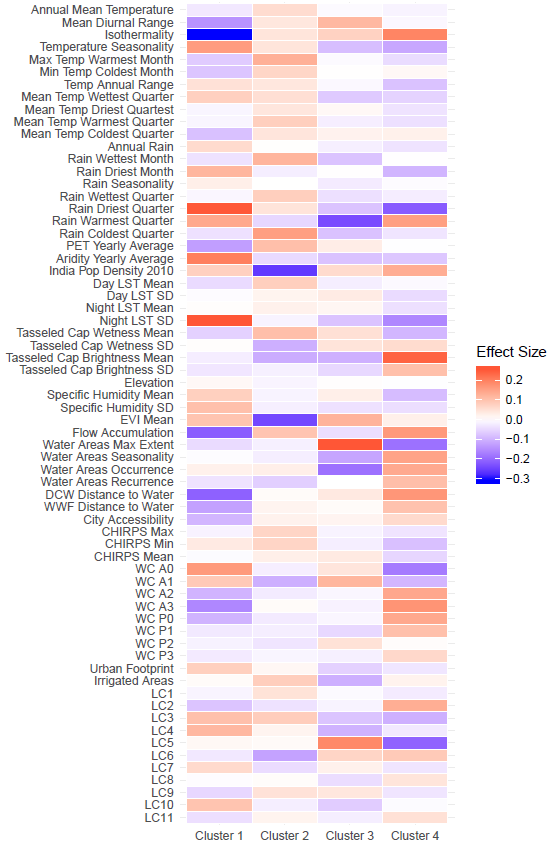
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Uninformative Prior** | | | |
| **Informative Prior** | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** |
| **Cluster 1** | 64 | 0 | 3 | 0 |
| **Cluster 2** | 0 | 72 | 0 | 11 |
| **Cluster 3** | 2 | 0 | 42 | 0 |
| **Cluster 4** | 21 | 4 | 14 | 39 |

**Predictive Performance was 0.51 😊**

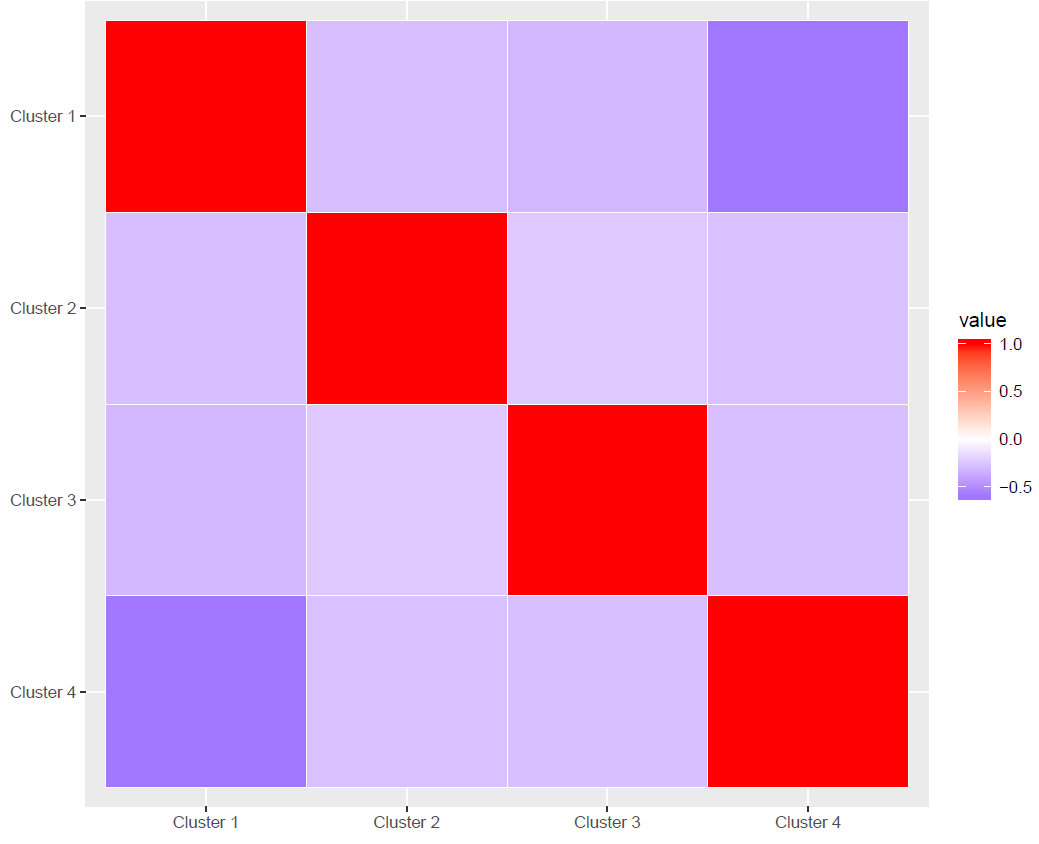
**Supplementary Figure 5: Uninformative Prior Results**



**Supplementary Figure 6: Species Coefficient Correlations Across Clusters**



**Supplementary Figure 7: Environmental Covariates Cluster Specific Coefficient Values**



**Supplementary Figure 8: Cross-Cluster Correlation For Ecological Covariates**

**References**

1. Dev, V. & Sharma, V. P. The Dominant Mosquito Vectors of Human Malaria in India. in *Anopheles mosquitoes - New insights into malaria vectors* (InTech, 2013). doi:10.5772/55215

2. Singh, N., Mishra, A. K., Chand, S. K. & Sharma, V. P. D Population dynamics of Anopheles culicifacies and malaria in the tribal area of Central India. *J. Am. Mosq. Control Assoc.* **15**, 283–290 (1999).

3. Kumari, S. *et al.* D VECTORIAL ROLE OF ANOPHELES SUBPICTUS GRASSI AND ANOPHELES CULICIFACIES GILES IN ANGUL DISTRICT, ORISSA, INDIA. *Southeast Asian J. Trop. Med. Public Health* **40**, 713–719 (2009).

4. Dutta, P., Bhattacharyya, D. R., Khan, S. A., Sharma, C. K. & Mahanta, J. Feeding patterns of Anopheles dirus, the major vector of forest malaria in north east India. *Southeast Asian J. Trop. Med. Public Health* **27**, 378–81 (1996).

5. Prakash, A., Bhattacharyya, D. R., Mohapatra, P. K. & Mahanta, J. Seasonal prevalence of Anopheles dirus and malaria transmission in a forest fringed village of Assam, India. *Indian J. Malariol.* **34**, 117–125 (1997).

6. Nanda, N. *et al.* Prevalence and incrimination of Anopheles fluviatilis species S (Diptera: Culicidae) in a malaria endemic forest area of Chhattisgarh state, central India. *Parasit. Vectors* **5**, 215 (2012).

7. Tripathy, A. *et al.* Distribution of sibling species of Anopheles culicifacies s.l. and Anopheles fluviatilis s.l. and their vectorial capacity in eight different malaria endemic districts of Orissa, India. *Mem. Inst. Oswaldo Cruz* **105**, 981–987 (2010).

8. Dev, V. & Manguin, S. D Biology, distribution and control of Anopheles (Cellia) minimus in the context of malaria transmission in northeastern India. *Parasit. Vectors* **9**, (2016).

9. Dev, V. D Anopheles minimus: Its bionomics and role in the transmission of malaria in Assam, India. *Bull. World Health Organ.* **74**, 61–66 (1996).

10. Korgaonkar, N. S., Kumar, A., Yadav, R. S., Kabadi, D. & Dash, A. P. Mosquito biting activity on humans &amp; detection of Plasmodium falciparum infection in Anopheles stephensi in Goa, India. *Indian J. Med. Res.* **135**, 120–6 (2012).

11. Kumar, A. *et al.* Anopheles subpictus carry human malaria parasites in an urban area of Western India and may facilitate perennial malaria transmission. *Malar. J.* **15**, (2016).

12. Gimnig, J. E., Ombok, M., Kamau, L. & Hawley, W. A. Characteristics of larval anopheline (Diptera: Culicidae) habitats in Western Kenya. *J. Med. Entomol.* **38**, 282–8 (2001).

13. Mattah, P. A. D. *et al.* Diversity in breeding sites and distribution of Anopheles mosquitoes in selected urban areas of southern Ghana. *Parasit. Vectors* **10**, 25 (2017).

14. Kumar Singh, R., Kumar, G. & Pradeep Kumar, Mittal Chand Dhiman, R. U Bionomics and vector potential of Anopheles subpictus as a malaria vector in India: An overview. *Int. J. Mosq. Res.* **1**, 29–37 (2014).

15. Amerasinghe, F., Indrajith, N., Sci, T. A.-C. J. & 1995, undefined. Physico-chemical characteristics of mosquito breeding habitats in an irrigation development area in Sri Lanka. *dl.nsf.ac.lk*

16. Weiss, D. J. *et al.* An effective approach for gap-filling continental scale remotely sensed time-series. *ISPRS J. Photogramm. Remote Sens.* **98**, 106–118 (2014).

17. Sinka, M. E. *et al.* The dominant Anopheles vectors of human malaria in the Asia-Pacific region: occurrence data, distribution maps and bionomic précis. *Parasit. Vectors* **4**, 89 (2011).

18. Funk, C. *et al.* The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Sci. Data* **2**, 150066 (2015).

also known as the kernel. This kernel is a positive-definite function of two inputs, and that defines the covariance matrix of our Gaussian Process. Specifically, these models utilise this kernel to define the prior covariance between any two data points, such that: