A climate-driven and field data-assimilated population dynamics model of sand flies

Kamil Erguler¹, Irene Pontiki¹, George Zittis¹, Yiannis Proestos¹, Vasiliki Christodoulou², Nikolaos Tsirigotakis², Maria Antoniou², Ozge Erisoz Kasap³, Bulent Alten³ & Jos Lelieveld⁴,¹

Sand flies are responsible for the transmission of leishmaniasis, a neglected tropical disease claiming more than 50,000 lives annually. Leishmaniasis is an emerging health risk in tropical and Mediterranean countries as well as temperate regions in North America and Europe. There is an increasing demand for predicting population dynamics and spreading of sand flies to support management and control, yet phenotypic diversity and complex environmental dependence hamper model development. Here, we present the principles for developing predictive species-specific population dynamics models for important disease vectors. Based on these principles, we developed a sand fly population dynamics model with a generic structure where model parameters are inferred using a surveillance dataset collected from Greece and Cyprus. The model incorporates distinct life stages and explicit dependence on a carefully selected set of environmental variables. The model successfully replicates the observations and demonstrates high predictive capacity on the validation dataset from Turkey. The surveillance datasets inform about biological processes, even in the absence of laboratory experiments. Our findings suggest that the methodology can be applied to other vector species to predict abundance, control dispersion, and help to manage the global burden of vector-borne diseases.

Leishmaniasis is a neglected tropical disease caused by protozoan flagellates vectored by phlebotomine sand flies. The disease is associated with about 20 Leishmania species, and it manifests itself in various forms (the leishmaniases) including zoonotic and anthropogenic visceral and cutaneous leishmaniasis. Primary vectors are Phlebotomus spp. in Eurasia and Lutzomyia spp. in the Americas, and a diverse zoonotic reservoir including rodents and domestic dogs has been described. Historically, leishmaniasis has been associated with poor living conditions and insufficient health-care; however, crowding and poor urban planning in combination with warming trends in global climate are rapidly exacerbating conditions that lead sand flies and leishmaniasis towards temperate territories.

The diseases caused by the Leishmania parasite are dynamic because sand flies are dependent on environmental, demographic and human behavioural factors. Such factors bring about changes in the habitat not only of the vectors but also of their natural hosts. At the same time, conditions resulting in immunosuppression in humans (like infection with human immunodeficiency virus (HIV) or organ transplantation-associated therapies) as well as the consequences of armed conflict, result in important changes in the epidemiology of the disease. In order to safeguard public health, it is necessary to take account of sand fly and leishmaniasis risk factors.

Progress in mathematical modelling of vector-parasite systems is limited by the complex epidemiology and the availability of reliable surveillance data. Nevertheless, known vector presence combined with climatic variables have been exploited for large-scale projections of habitat suitability. Fischer et al. performed environmental niche modelling with the maximum entropy methodology (MaxEnt) to model habitat suitability and project routes of dispersal across Central Europe driven by climate change. Pigott et al. used boosted regression trees to infer environmental constraints and project global habitat suitability. In a more recent study, Meneguzzi et al. performed environmental niche modelling to associate disease vectors with cutaneous leishmaniasis incidents in Southeastern Brazil.

¹Energy, Environment and Water Research Center, The Cyprus Institute, 2121, Aglantzia, Nicosia, Cyprus. ²Laboratory of Clinical Bacteriology, Parasitology, Zoonoses and Geographical Medicine, School of Medicine, University of Crete, Heraklion, Crete, Greece. ³Faculty of Science, Department of Biology, Ecology Section, Hacettepe University, 06800, Beytepe-Ankara, Turkey. ⁴Department of Atmospheric Chemistry, Max Planck Institute for Chemistry, D-55128, Mainz, Germany. Correspondence and requests for materials should be addressed to K.E. (email: k.erguler@cyi.ac.cy)
While environmental niche modelling provides a large-scale overview of possible environmental limitations of the vector species, mechanistic modelling with explicit links to the underlying biological processes offers a more detailed understanding of the environmental dependence of vector biology and disease epidemiology. Oshaghi et al. developed a population dynamics model of *Ph. papatasi* driven by local meteorological station data in Iran. The authors used temperature thresholds from the literature to estimate accumulated degree days for life history processes, and adjusted laboratory derived parameters for field populations. The resulting correlation between predictions and observed population abundance suggests that laboratory-informed population dynamics modelling could be predictive provided that the parameters are properly adjusted for field conditions.

One of the key limitations of most population dynamics and disease transmission models, including the ones proposed for sand flies and leishmaniasis, is their direct exploitation of laboratory experiments on key life-history traits, such as survival and development. Parameters derived from controlled experimental conditions may not be readily applicable for field populations especially while microclimate conditions around potential breeding sites cannot be reliably predicted with existing computational capacity and climate models. For these reasons, we recently proposed a Bayesian approach to develop population dynamics and disease transmission models by combining experimental data with field observations and observing the dynamics under the influence of large-scale environmental variables. The emerging models demonstrated wide applicability and high predictive ability. However, the amount of experimental data on a vector or disease of interest is often limited due to the small number of experimental studies on model systems, the variety of vector and pathogen species, and the lack of host specificity.

Here, we propose an adaptation of our Bayesian approach to assess the information content of field data obtained from multiple locations in informing predictive population dynamics models. We focus on three important vector species: *Phlebotomus neglectus*, *Phlebotomus tobbi* and *Phlebotomus papatasi*. While *Ph. neglectus* and *Ph. tobbi* are potent vectors of *Leishmania infantum*, which causes anthroponotic and zoonotic visceral leishmaniasis, *Ph. papatasi* is a vector of *Leishmania major*, which causes zoonotic cutaneous leishmaniasis. We developed a flexible and extensible framework for developing population dynamics models from field observations and performing climate-driven projections at a regional or global scale. We demonstrate that our approach is easily adjustable to study multiple vector species based on available surveillance data.

**Methods**

We developed a generic discrete-time stochastic model to represent the environmental dependence of sand fly populations. We present a schematic representation of the model in Fig. 1, and list model parameters in Table 1 together with domain boundaries used for parameter inference (see sec. Vector dynamics model). We implemented the model in ANSI C and integrated it with the *albopictus* Python package (v.1.6), which is available from the Python package index.

**Vector dynamics model.** We designed a stochastic difference equations model with Markov property, which implies that the state of a population on a given day is affected only by the state of the population and the environmental conditions on the previous day. We defined the population with the number of sand flies and the age structure in each stage. A single iteration comprises the cumulative effect of all stage transformations taking place during a day. Transformations include daily survival, development, egg laying, and, in case of an ongoing surveillance activity, being captured. For each transformation, we defined daily probabilities such as the daily probability of death or egg hatching.

Life history parameters of sand flies depend mainly on temperature, and, to a certain extent, on humidity. In an attempt to standardise environmental dependency for the extensively diverse *Phlebotomus* genera, we employed single- or double-sided sigmoid curves relating temperature and humidity to survival and development. Sigmoid curves are frequently used in machine learning and fuzzy-logic classification where class
Table 1. Model parameters with reference intervals.

| Parameter | Definition | Prior interval | Notes |
|-----------|------------|----------------|-------|
| \(n_0(0), n_i(0), n_0(i), n_i(i)\) | Initial number of eggs, larvae, pupae, and adults, respectively | \(U(0, 10^4)\) | |
| \(d_{E2}, d_{L2}, d_{P2}\) | Minimum mean development time of eggs, larvae, and pupae, respectively | \(U(1, 50)\) days | |
| \(d_{E3}, d_{L3}, d_{P3}\) | Maximum mean development time of eggs, larvae, and pupae, respectively | \(U(1, 50)\) days | Must be higher than the minimum |
| \(d_{E1}, d_{L1}, d_{P1}\) | Median developmental temperature for eggs, larvae, and pupae, respectively | \(U(-10, 50)\)°C | |
| \(d_{A}\) | Minimum gonotrophic cycle duration | \(U(1, 50)\) days | |
| \(d_{A1}\) | Median temperature for gonotrophic cycle duration | \(U(-10, 50)\)°C | Must be higher than the minimum |
| \(P_{E1}, P_{L1}, P_{P1}, P_{A1}\) | Maximum daily survival probability of eggs, larvae, pupae, and adults respectively | \(U(0, 0.999)\) | |
| \(P_{E2}, P_{L2}, P_{P2}, P_{A2}\) | Minimum survival temperature threshold for eggs, larvae, pupae, and adults respectively | \(U(-10, 50)\)°C | |
| \(P_{E3}, P_{L3}, P_{P3}, P_{A3}\) | Maximum survival temperature threshold for eggs, larvae, pupae, and adults respectively | \(U(-10, 50)\)°C | Must be higher than the minimum |
| \(E_0\) | Maximum number of eggs laid in a single gonotrophic cycle | \(U(0, 100)\) | |
| \(F_{A}\) | Minimum temperature threshold for oviposition | \(U(-10, 50)\)°C | |
| \(F_{A1}\) | Maximum temperature threshold for oviposition | \(U(-10, 50)\)°C | Must be higher than the minimum |
| \(p_o\) | Probability of surviving oviposition | \(U(0, 0.5)\) | |
| \(p_f\) | Probability that a female adult emerges from a larva | \(U(0, 1)\) | |
| \(p_{dev}\) | Daily probability of getting caught | \(U(10^{-4}, 10^{-1})\) | 0.01–10% coverage is assumed |
| \(\Psi_{T}\) | Temperature threshold for dormancy | \(U(-10, 50)\)°C | |
| \(\Psi_{RH}\) | Relative humidity threshold for egg and larva survival | \(U(0, 1)\) | |
| \(\Psi_{E}\) | Reduction in larva and pupa development probability and adult egg laying due to limited habitability of a breeding site | \(U(0, 1)\) | |

Boundaries are poorly defined^{26}. Here, we employed sigmoid curves as generic approximations of environmental dependence where experimental observations are limited and observational error is large.

In order to limit complexity, we took account of the scarcity of experimental data, and assumed that temperature and humidity affect survival independently. We assumed that temperature strongly drives development and survival of all stages, and humidity is a requirement for the survival of the early immature stages, i.e. eggs and larvae. In essence, sufficient humidity is required for eggs and larvae to sustain viability^{27}, and when humidity is sufficiently high, survival becomes dependent mainly on temperature. We used a single-sided sigmoid curve to describe egg survival as a function of relative humidity,

\[ P_{\text{surv}_{E}} = 1 \left(1 + e^{\Psi_{RH} - RH} \right)^{-1} \]

where RH represents daily average relative humidity and \(\Psi_{RH}\) represents the minimum RH for survival. The normalisation factor, 0.01, is to constrain the left and right tails of the curve within 0–1.

We used a double-sided sigmoid curve to describe temperature-driven survival. For instance, we described the daily probability of temperature-driven egg survival with

\[ P_{\text{surv}_{E}} = \frac{P_{E1}}{1 + e^{T - T_{EE}}} \]

where \(T\) is daily average near-surface temperature (2 meters above ground). The parameters constraining the curve are defined in Table 1.

Survival probability changes daily as a result of temperature and humidity differences. In a hypothetical scenario where these variables are fixed, the survival process can be described with a geometric distribution with a constant probability \(p = 1 - P_{\text{surv}}\). In this case, the average life span can be calculated as \((1 - p)/p\).

We assumed that the development process is independent of survival, and modelled development time for eggs, larvae, pupae, and the gonotrophic cycle duration for adult females with a gamma distribution. Under constant environmental conditions, we assumed that each process lasts on average \(\mu\) days with a standard deviation of \(\sigma\). Together, \(\mu\) and \(\sigma\) describe the probability of the duration of a development process. The resulting daily probability for the completion of development, \(P_{\text{dev}}\), can be written as

\[ P_{\text{dev}} = \frac{\Pr(d < x \leq d + 1)}{\Pr(d < x)} \]
for day $d$, where $x \sim \Gamma(\mu, \sigma)$. In order to limit the total number of parameters for this generic model structure, we assumed that the variance of the gamma distribution is equal to its mean ($\sigma^2 = \mu$). We argue that this assumption is reasonable with regards to the experimental data (see sec. Experimental datasets); however, improvements and customisations are advisable on specific life history traits as more data become available.

Immature stage development and the gonotrophic cycle are hampered at extremely low temperatures$^{1,24}$. In order to model these processes, we used a single-sided sigmoid curve,

$$d_{\text{mean}} = d_{\text{max}} - \frac{d_{\text{max}} - d_{\text{min}}}{1 + e^{-(T - T_{\text{med}}) / \sigma}}$$

where $d_{\text{min}}$ is the minimum, $d_{\text{max}}$ is the maximum duration, and $T_{\text{med}}$ is the temperature with the median duration. Equation 1 results in a mirrored S-shaped curve yielding long development times at low temperatures and a rapid development process at high temperatures.

Although female sand flies usually require blood feeding to develop eggs, egg laying and feeding patterns are discordant (egg laying can be delayed until after multiple blood meals) in certain species$^{19}$, and reproduction is autogenous (egg laying requires no blood meal) in others$^{22}$. Therefore, we assumed that the duration of the gonotrophic cycle is influenced mainly by temperature$^{24}$. Mortality is strongly elevated after oviposition, where only a few females survive for the second, probably last, oviposition$^{28}$. In order to model this unusually high death rate, we defined $p_{\delta}$ as the probability of surviving egg laying for each female. At each gonotrophic cycle, a healthy female can lay up to 100 eggs$^{29}$, which is observed to depend on temperature$^{24}$. According to this, we used a double-sigmoid curve for the number of eggs laid,

$$P_{\text{EA}_t} = \frac{F_{A_1}}{(1 + e^{(t - T_0) / \tau}) (1 + e^{(T_1 - t) / \tau})},$$

where $F_{A_1}$ is maximum number of eggs a female can lay per cycle, and $F_{A_1}$ and $F_{A_1}$ are the maximum and minimum, respectively, temperatures for fecundity.

Aside from egg laying, we assumed that male and female sand flies share similar survival and developmental characteristics in all life stages$^{30,32}$. In order to determine the number of fertile female adults, we defined $p_{\Phi}$ as the probability of emergence of a female adult sand fly from a developing pupa.

Facultative diapause, a mechanism of sustaining development in response to environmental factors, is reported in the 4th instar larva in several species including Ph. neglectus, Ph. tobbi, and Ph. papatas$^{10,19,31}$. In order to accommodate the state of environmentally driven dormancy, we defined $\Psi_d$ and updated the development time given in Eqn. 1. According to this, when temperature is lower than this threshold, $T < \Psi_d$, larval development time is set to an extreme value (1000 days) to delay for as long as possible. Otherwise, development proceeds as given in Eqn. 1.

Sand fly abundance and diversity are affected strongly by the availability of appropriate breeding grounds. Although precipitation is not a direct influence, existence of rocky slopes, vegetation, availability of decaying organic matter, and an alkaline soil pH are known to contribute to abundance and diversity$^{32-34}$. In order to capture breeding site conditions, we defined $\Psi_d$ as a location-specific fraction, which reduces fecundity, prolongs egg laying interval, and extends the development times of larvae and pupae. According to this, $\Psi_d = 1$ indicates perfect breeding conditions where development and fecundity depend only on temperature and humidity. When $\Psi_d < 1$, development rate is reduced due to adverse conditions such as limited availability of food or resting sites. We implemented this effect by increasing average development times ($\mu$ and $\sigma$) and reducing fecundity ($P_{\text{EA}_t}$) by a factor of $\Psi_d$. We note that $\Psi_d$ can only be inferred when performing inference on multiple locations. When a single location is in focus, $\Psi_d$ merges with the rest of the parameters that drive development rate, e.g. $P_{\Phi}$, $P_{\delta}$, $P_{\delta}$, and $P_{\delta}$.

The constant indicator implies that a breeding site provides sustained levels of availability and habitability. In addition, we assumed that the main drivers of population dynamics are climatic variables, and that population growth under the influence of these drivers would not reach high levels at which density or predation becomes a limiting factor. The effect of breeding site conditions on survival, therefore, is indirect as delayed development increases exposure to adverse weather conditions. In the current form of the model, $\Psi_d$ is an arbitrary indicator and a first-order approximation of a complex and time-dependent phenomenon. Although this might be adequate for a generic model, it is subject of improvement in future work.

Although the flight range of adult sand flies is limited, they are known to be dispersed over long distances with the help of wind$^{27,31}$. We postpone the discussion of the effect of migration and dispersion to future work where spatiotemporal modelling will be considered. In the present context, we assumed that migration and wind have negligible effects on the abundance of sand fly populations.

In addition to environmental constraints, we modelled the impact of surveillance activities on the abundance of adult sand flies. We defined $P_{\delta}$ as the daily probability of either a male or a female adult being captured and removed from the system. Accordingly, we defined the probability of an observation ($\delta$) as the product of the binomial capture probabilities at each time point,

$$f(\delta|y) = \prod_{t=1}^{T} \text{Binom}(y(t), \delta(t), p_{\delta}) \quad \text{for} \quad y(t) \geq \delta(t) \quad \forall t$$

$$0 \quad \text{otherwise}$$
where \( y(t) \) is the predicted number of adult males or females, and \( \delta(t) \) is the number of captured sand flies for day \( t \). Predictions and observations comprise the number of female or male sand flies collected over three days as given in the surveillance procedure (see sec. Surveillance and environmental datasets below). In this equation, \( \text{Binom}(n, m, p) \) represents the binomial probability of choosing \( m \) sand flies out of a total of \( n \) with probability \( p \).

We note that \( y(t) \) needs to be greater than or equal to \( \delta(t) \) for all time points for the binomial probability to be defined. We handled the special case when this condition is not satisfied by replacing the binomial probability with an infinitesimal indicator value (\(< e^{-1} \)) at problematic time points. The non-zero probability provides an indication for the degree of mismatch and a gradient to guide the optimisation routine towards more plausible parameter values. We note the fraction of times simulations yield such problematic time points and indicate the degree of success (\( \rho \)) throughout the text.

Likelihood of an observation in the context of Bayesian inference can be calculated as the weighted average of the binomial capture probability,

\[
f(\delta|\theta) = \int f(\delta|y)f(y|\theta)dy, \tag{2}
\]

where \( y \) is a prediction, i.e. a sample, from the model when its parameter values are given by \( \theta \). Due to limited computational resources, we approximated the likelihood with a Monte Carlo approximation,

\[
f(\delta|\theta) \approx \frac{1}{N} \sum_{i=1}^{N} f(\delta|y_i)f(y_i|\theta), \tag{3}
\]

where \( N \) is the number of random samples from the model used to approximate the likelihood. More accurate approximations can be derived by using a larger sample size at the expense of computational resources; however, we found that \( N = 3 \) performs well in identifying an optimum fit (see Results and Discussion).

The model captures the essence of the environmental dependency of sand fly populations; however, we expect that species-specific differences, microclimate conditions around the breeding sites, day-night differences, intrinsic variability associated with the limited population size, and many other unaccounted factors may contribute to deviations between simulations and observations. Inevitably, such differences yield a very small likelihood value. In order to facilitate working with small values, we opted to work with the log-transformed likelihood, which we designated as the score,

\[
S(\delta|\theta) = -\ln \frac{1}{3} \sum_{i=1}^{3} f(\delta|y_i)f(y_i|\theta). \tag{4}
\]

**Inference and posterior sampling**

Having defined an objective function (Eqn. 4), we employed a uniform prior distribution as presented in Table 1 (\( f(y|\theta) = c \)), performed parameter optimisation, and obtained posterior samples of \( \theta \) using the python (v2.7) implementation of the hoppMCMC (v0.6) algorithm\(^{17,18}\). The hoppMCMC algorithm uses an adaptive Markov chain Monte Carlo (MCMC) method to obtain samples from the posterior distribution\(^{35}\).

For sufficiently large \( N \), the Monte Carlo approximation (Eqn. 3) converges to the true likelihood, and the algorithm yields accurate posterior samples. When \( N \) is small, however, the error in the estimate, which is proportional to \( 1/\sqrt{N} \), may result in a prohibitively low acceptance rate. In order to compensate for small sample size, we performed posterior sampling with a tolerance. We scaled down the likelihood estimate exponentially with a constant, \( C < 1 \), to ensure that a minimum acceptance rate of 1% can be achieved. The constant corresponds to the inverse of the annealing temperature of the hoppMCMC algorithm. We report \( C \) where applicable. Details of inference and posterior sampling can be found in Supplementary File S1.

When reporting inference results, we adopted the concept of posterior modes as reported in previous work\(^{17,18}\). In essence, we assumed that the posterior distribution comprises a set of local maxima analogous to the local minima frequently encountered in optimisation. We assumed that the locality of each maximum leads to similar model behaviour, and the extent of each locality is dictated by the sensitivity around the maxima. In case of high sensitivity, only the parameter values in the vicinity of the maximum yield similar behaviour, and the behaviour changes drastically at greater distance. Although it is not trivial to sample from the entire posterior distribution, it is relatively straightforward to sample from the vicinity of a local maximum by starting a Markov chain from around that point\(^{17}\). Posterior samples obtained as such are labelled as samples from the same posterior mode. Throughout the text, we report posterior modes (single or multiple) as possible explanations for the observations.

**Experimental Datasets**

Due to the cost and effort required to reliably determine life history parameters, comprehensive information on many sand fly species is yet to be obtained. Kasap et al. compiled a reliable data source on Ph. papatasii in two reports: Kasap et al.\(^{11}\) and Kasap et al.\(^{24}\). Here, we derived daily survival probabilities, development times, and fecundity for a range of temperatures using these data (Fig. 2).

We emphasise that these parameters originate from controlled environmental conditions, which include temperature, humidity, feeding, and photoperiod. Therefore, care must be taken when extrapolating them to be used with mathematical models due to pertaining structural restrictions (such as the existence of four distinct life stages) and simplifying assumptions (such as the use of daily average temperatures).
for survival, which corresponds to 1% of the range (0–1). In order to assess the

**Results and Discussion**

We assessed the inferability of a biologically-plausible environmentally driven population dynamics model using only sand fly surveillance data as the basis of inference. We assessed the predictive capacity of models inferred using various degrees of information, and presented guidelines for extending this procedure to other species.

**Multiple versions of the model are informed by the surveillance data.** The sand fly model comprises a total of 37 parameters including 6 that are location-specific (the four initial conditions, the degree of habitability, and the daily probability of capture — see sec. Vector dynamics model). We inferred all of these using only a single location for each of the three species: *Ph. papatasi*, *Ph. neglectus*, and *Ph. tobbi* (see sec. Inference and
posterior sampling). Then, we combined the three locations to infer all for *Ph. papatasi*, the only species found in all locations, while allowing three different values for the location-specific parameters.

As a result, when we used only one location for inference, we identified one major posterior mode for each. We identified multiple posterior modes when we combined all locations, and, here, we report three of the most representative and most probable ones labelled as “Combined A”, “Combined B”, and “Combined C”. In Table 3, we report the annealing temperature ($C$), the degree of success ($\rho$), the average score ($\mathbb{E}(\delta\theta|\theta)$), and the absolute value of the determinant of the covariance matrix ($\mathbb{U}$). We report $\mathbb{U}$ as a measure of dispersion in posterior modes.

The likelihood function is proportional to the size of a dataset and the number of sand flies collected (Eqn. 2); therefore, a large number of collections or more data points yield high likelihood values. In turn, this constrains the extent of a posterior mode. In agreement with this, Fodele (the largest dataset) yields a less dispersed posterior mode than Geri (the smallest dataset). Combined inferences yield highly constrained posterior modes due to the cumulative size of the combined dataset. We observed the lowest dispersion for *Ph. neglectus* which was collected in large numbers. In order to preserve 1% acceptance rate in sampling, a higher annealing temperature was needed for this dataset.

Estimates of kernel density of the scores associated with the inference of *Ph. papatasi* can be seen in Fig. 4. We found that single location inferences explain each location better than combined inferences. Posterior modes labelled “Combined A” and “Combined B” yield balanced fits to the three locations, while “Combined C” tends to fit Fodele better than the other two locations. Resulting matches with the data for all inferences are shown in Supplementary Fig. S1.

In Fig. 5, we present the match between surveillance data and model predictions with Combined A. We note that the median predictions for winter, despite being explicitly set to zero in the training set, are relatively low,
but rarely zero in Fodele. This suggests that a continuous, yet minimal, sand fly activity might be present in the region during winter. One of the consistent predictions by all posterior modes (Supplementary Fig. S1) is that the beginning of the sand fly high season may be several months earlier in Fodele than the beginning of the surveillance period (April).

An interesting pattern in the surveillance data from Fodele is a peak occurrence both in males and females in June in both years of observations. Neither the single-location inference nor the combined ones fully explain these peaks for both sexes. The mismatch is indicative of a biological mechanism not accounted for in this version of the model. This could be, for instance, an alternative type of dormancy, such as obligate diapause, where termination is spontaneous and independent of external stimuli. Seasonal changes in breeding sites or feeding patterns could also result in the observed peaks. The understanding of this mechanism requires further iterations of the model, which will be the subject of continued research.

**Surveillance data sufficiently informs predictive models.** In this context, we constrained environmental dependence with model structure and longitudinal surveillance data. Although not included in the likelihood nor considered as a prior probability, we assessed whether laboratory-driven life history parameters can be inferred correctly for *Ph. papatasi*. In Table 4, we display the average matching scores (\(\Pi(\theta)\)) and their standard deviations for each posterior mode. In Supplementary Fig. S2, we plot all functional forms of environmental dependences included in the model, and in Fig. 6, four discerning dependences inferred for Fodele where the posterior modes are the least dispersed.

We found that combined inferences that fit each location equally (Combined A and B) closely resemble the environmental dependence determined with laboratory experiments. The analysis indicated that although they yield more constrained posterior modes, larger datasets from single locations do not necessarily yield better predictions of environmental dependence. In addition, the unbalanced fit among multi-location inferences, Combined C, which leans towards Fodele, does not agree well with laboratory data.

In the framework of this study, we obtained the best-possible agreement by omitting the likelihood and performed inference only on the prior probability. When we compared this best fit with the agreement scores (\(\Pi(\theta)\)) of Combined A and B, we found that although Combined A and B demonstrate a close match, they do not overlap with the best fit (Fig. 7). The evident divergence from the maximum agreement indicates that a certain degree of deviation from the experimentally-derived values is needed in order to explain the surveillance data. Accordingly, we argue that direct use of experimental results in mathematical models of population dynamics might be misleading.

In order to assess the information content of the surveillance data, we performed parameter sensitivity analysis using a multivariate Gaussian approximation around the maximum of each posterior mode. We calculated the sensitivity matrix \(\mathcal{H}\) as the inverse of the variation of each Gaussian, and normalised with the range of the parameter. In Table 3, we display the associated sand fly species, annealing temperature (\(C\)), degree of success (\(\rho\)), average and standard deviation of the score (\(S(\theta|\theta)\)), and the extent of dispersion in posterior modes (\(\mathcal{U}\)).

| Species | Type | Area | \(C\) | \(\rho\) | \(S(\theta|\theta)\) | \(\mathcal{U}\) |
|---------|------|------|------|------|----------------|-------|
| *Ph. neglectus* | Single | Fodele | 4 | 100.0% | 733.53 ± 19.12 | 79.6714 |
| *Ph. tobbi* | Single | Steni | 2 | 100.0% | 44.29 ± 5.49 | 147.879 |
| Combined A | Geri | 2 | 100.0% | 33.76 ± 1.35 | 104.499 |
| Combined B | Steni | 2 | 100.0% | 54.61 ± 7.65 | 106.085 |
| Combined C | Fodele | 2 | 100.0% | 44.34 ± 3.63 | 103.678 |

Table 3. Posterior modes obtained using the training datasets. Single or combined inferences are listed with the associated sand fly species, annealing temperature (\(C\)), degree of success (\(\rho\)), average and standard deviation of the score (\(S(\theta|\theta)\)), and the extent of dispersion in posterior modes (\(\mathcal{U}\)).
daily survival probability of adults ($p_{\text{A}_t}$), On the other hand, the relative humidity threshold ($\Psi_{\text{RH}}$), the temperature threshold for dormancy ($\Psi_{\text{T}}$), and initial conditions do not have a high impact on model dynamics.

High sensitivity is an a priori expectation for $p_{\text{A}_t}$, as it is strongly informed by the presence of both female and male sand fly counts in surveillance datasets. High sensitivity was also expected for $\Psi_{\text{RH}}$, being directly linked with multiple biological processes such as larva and pupa development, and fecundity. Lastly, high sensitivity for $p_{\text{A}_t}$ was expected due to the direct effect on the observed number of sand flies.

We note that the sensitivity measure we employed depends largely on the characterisation of the posterior distribution. In this context, sensitivity is a measure of impact of a parameter on the dynamics of male and female sand flies. Therefore, the low impact parameters, such as temperature and relative humidity threshold for dormancy, and initial conditions are relatively less potent in determining adult sand fly dynamics according to most of the posterior modes. On the other hand, parameters controlling pupa development and adult survival have high sensitivities due to being directly related to adult population size. An interesting finding is the high sensitivity of the gonotrophic cycle duration that is likely due to the negative effect of egg laying on the survival of adult sand flies.

**Multi-location inferences yield predictive models.** We tested model predictability over the surveillance dataset from Adana. The test set contains sightings of Ph. papatasi, Ph. tobbi, and the Ph. major s.l. species complex (corresponding to either Ph. neglectus or Ph. syriacus in Adana) for the two consecutive years 2011–2012 (see sec. Surveillance and environmental datasets). As a result, we found that the single-location inferences and Combined C do not yield useful predictions for any of the species (see Supplementary Fig. S3). The majority of model predictions using these posterior modes comprise very few or no comparable trajectories with the observations (Table 5). Inclusion of laboratory data as a prior probability for single-location inferences elevated the success rate only for the case of Ph. papatasi.

We found that Combined A and B — the multi-location inferences that demonstrate a balanced fit to each location — successfully predict Ph. papatasi abundance (Fig. 9). We obtained high variability in the predictions due to the intrinsic stochasticity of the model and the uncertainty in parameter values. Despite this, we calculated high success rates. We followed the single-location inference procedure to identify an optimum posterior mode for the Ph. papatasi in the test set. We found that the median trajectories with Combined A and B are close to the median of the optimum posterior mode endorsing the reliability of the predictions.

We note that Combined A and B each have different variations for the three locations in the training set. It appears that the main difference between each variation is the overall vector abundance, which is mainly controlled by breeding site habitability, $\Psi_{\text{B}}$ (Fig. 10). Despite the differences between Combined A and B in the remaining parameters, we observed that the variations with $\Psi_{\text{B}}$ between 3.5 and 4.5 explain better the observed abundance in the test set.
Conclusion

Comprehensive studies concerning the environmental dependence of vector species require considerable effort and the availability of suitable rearing facilities. For this reason, vital information required for predictive modeling is scarce for many important vectors. It is clear that flexible strategies exploiting already available or relatively easily obtainable data are needed. Vector surveillance, which is usually performed by authorities for monitoring and control, provides a viable source of information.

Here, we demonstrate that predictive models can be constructed by using field observations, while laboratory studies inform model structure and impose known biological constraints. The model we presented is structurally informed by the known biological mechanism and the environment-dependence of well-studied sand fly species. Parameters of the model are configured to represent a specific vector species by exploiting various amounts of surveillance data.

Figure 5. Agreement between *Ph. papatasi* surveillance data and model predictions with Combined A. Number of adult sand flies collected from Geri, Steni, and Fodele are plotted as bars in (a–c), respectively. Solid lines and shaded areas indicate the median and the 95% range of the predictions. Female and male adult sand flies are marked in red and blue, respectively.
Our analyses corroborate that in order to construct predictive models, surveillance data collected from multiple independent sites are needed. In addition, a posterior sample of parameters should demonstrate a balanced fit by representing each site adequately to yield a predictive model. As a result, we observed that *Ph. papatasi* dynamics in Adana, Turkey, can be predicted with parameters inferred using surveillance data from Crete and Cyprus.

We note that the surveillance areas considered in this study share many environmental characteristics. Although our results suggest that the model may be applicable to other regions with similar characteristics,
large-scale regional and global applications should take account of diverse surveillance datasets comprising observations from different environmental conditions.

Our results suggest that the most predictive parameter configurations are also informative about the environmental dependence of *Ph. papatasi*. Although laboratory-derived life-history parameters were not explicitly imposed in the form of prior probability, the inferred parameter values agree well with the experimentally-determined values. Regardless of the agreement, we obtained large variations especially in the lower and upper bounds of temperature dependence. This is indicative of the absence of observations during such extreme conditions, which can be compensated by constraining the corresponding parameters explicitly with regards to expert knowledge.

We note that humidity, despite being indicated as a key environmental dependence, was not informed adequately by the surveillance data — only a substantially low threshold was identified as critical. It is possible that a more biologically-tailored model structure and more detailed observational datasets are needed to provide the required resolution to discern humidity dependence. We suggest the inclusion of day-and-night temperature

---

**Table 5.** Model validation over the test set from Adana.

| Species    | Type | Area       | $\rho$ | $S(\theta)$             |
|------------|------|------------|--------|--------------------------|
| *Ph. major s.l.* | Single | Fodele | 0      | —                        |
| *Ph. tobbi*    | Single | Steni   | 17.5%  | 1358.76 ± 1273.42        |
|               | Single | Geri    | 0      | —                        |
|               | Single | Steni   | 0.6%   | 1252.01 ± 153.29         |
|               | Single | Fodele | 0.7%   | 4423.84 ± 161.76         |
| *Ph. papatasi* | Combined A | Geri     | 0.2%   | 2221.41 ± 124.49         |
|               | Combined A | Steni   | 99.3%  | 12522.04 ± 12859.06      |
|               | Combined A | Fodele | 96.6%  | 744.00 ± 478.24          |
|               | Combined B | Geri    | 0      | —                        |
|               | Combined B | Steni   | 91.9%  | 5654.38 ± 7651.19        |
|               | Combined B | Fodele | 62.4%  | 1250.66 ± 740.88         |
|               | Combined C | Geri    | 0      | —                        |
|               | Combined C | Steni   | 0      | —                        |
|               | Combined C | Fodele | 0      | —                        |

---

**Figure 7.** Agreement of Combined A and B with laboratory-derived parameter values. Kernel density estimates of the parameter agreement scores ($\Pi(\theta)$) are plotted for the three versions of Combined A and B along with the best fit, which was obtained by excluding the likelihood. Frequency of the best fit distribution (black) is scaled down by a factor of 0.05 to aid in visualisation.

**Figure 8.** Parameter sensitivity analysis of posterior modes for *Ph. papatasi*. Color gradient shows the logarithm of the parameter sensitivities where darker colours indicate higher values. Parameters are sorted from left to right with a descending order with respect to the sum of their sensitivities in posterior modes Combined A and B.


and humidity differences as a possible next step since such data were reported to correlate strongly with vector abundance. We also note the possibility of the existence of a posterior mode with a higher probability and predictive capacity than the ones identified in this study. Further improvement of the model and the identification of predictive models for *Ph. neglectus* and *Ph. tobbi* will be the subject of future research.

An additional factor contributing to the variance between observations and predictions is the assumed range of applicability of our model. The distance between collection sites and the closest available meteorological stations is a limiting factor. Furthermore, it is not trivial to include microclimatic conditions due to the difficulty of obtaining or predicting such information. At present, the most detailed global climate simulations are provided at 0.25 degree resolution, which corresponds to about 25 km near the equator. With the improvement of global and regional climate models and the public availability of such data, we expect that the accuracy of vector and disease dynamics predictions will improve considerably.

It is important to note that none of the inferred parameter values, despite some demonstrating high predictive ability, provides the maximum possible fit to the laboratory-derived values. This indicates that extra care is needed when applying laboratory data to a larger scale. Experiments reflect vector biology under controlled environmental conditions. The scale of such studies may be related, to some extent, to the microclimatic environment around breeding sites. The scale of a typical field study, however, is several square kilometers at best. The methodology we presented offers a potential solution to this problem by allowing environmental dependence to be informed, and therefore adjusted to data collected in the filed.

**Data Availability**

All data generated or analysed during this study are included in this published article (and its Supplementary Information files).

**References**

1. Ready, P. D. Biology of Phlebotomine Sand Flies as Vectors of Disease Agents. *Annu. Rev. Entomol* 58, 227–50, https://doi.org/10.1146/annurev-ento-120811-153557 (2013).
2. Pigott, D. M. *et al.* Global distribution maps of the leishmaniases. *eLife* 3, 1–21, https://doi.org/10.7554/eLife.02851 (2014).
3. Desjeux, P. The increase in risk factors for leishmaniasis worldwide. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 95, 239–243, http://www.sciencedirect.com/science/article/pii/S0035920301902238, https://doi.org/10.1016/S0035-9203(01)90223-8 (2001).
42. Halimi, M., Cheghabaleki, Z. Z., Modrek, M. J. & Delavari, M. Temporal Dynamics of Phlebotomine Sand Flies Population in Response to Ambient Temperature Variation, Bam, Kerman Province of Iran. *Annals of Global Health* **82**, 824–831, https://doi.org/10.1016/j.aoghe.2016.07.009 (2016).

43. Sheffield, J., Goteti, G. & Wood, E. F. Development of a 50-year high-resolution global dataset of meteorological forcings for land surface modeling. *Journal of Climate* **19**, 3088–3111, https://doi.org/10.1175/JCLI3790.1 (2006).

Acknowledgements

We are indebted to Christoforos Pavlou and Apostolos Mazeris for their efforts on sand fly collection. KE would also like to thank L. Leticia Ramirez Ramirez for her valuable comments concerning the methodology. We acknowledge the E-OBS dataset from the EU-FP6 project ENSEMBLES (http://ensembles-eu.metoffice.com) and the data providers in the ECA&D project (http://www.ecad.eu). We acknowledge the Cy-Tera Project (NEW INFRASTRUCTURE/STRATEGIC/0308/31), which is co-funded by the European Regional Development Fund and the Republic of Cyprus through the Research Promotion Foundation, for providing the computational resources used in this research. We acknowledge Stamen Design (http://stamen.com, CC BY 3.0, http://creativecommons.org/licenses/by/3.0) for map tiles and OpenStreetMap contributors (http://www.openstreetmap.org/copyright) for map data. KE would like to acknowledge the European Union Scholarship Programme for the Turkish Cypriot Community for financial support. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

K.E. conceived the study, developed the model, performed the analyses, and wrote the paper. I.P. contributed to model development. G.Z. and Y.P. curated the environmental datasets. V.C. and N.T. contributed to surveillance data collection. M.A., O.E.K. and B.A. provided surveillance data and validated the results. J.L. supervised the study and validated the results. All authors contributed to the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-38994-w.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019