ECOGRAPHY

Research article

Improving distribution models of sparsely documented disease vectors by incorporating information on related species via joint modeling

Stacy Mowry, Sean Moore, Nicole L. Achee, Benedicte Fustec and T. Alex Perkins

University of Notre Dame, Notre Dame, IN, USA

Correspondence: Stacy Mowry (smowry@nd.edu)

Ecography
2024: e07253
doi: 10.1111/ecog.07253
Subject Editor: Dan Warren
Editor-in-Chief: Sydne Record
Accepted 28 March 2024

A necessary component of understanding vector-borne disease risk is accurate characterization of the distributions of their vectors. Species distribution models have been successfully applied to data-rich species but may produce inaccurate results for sparsely documented vectors. In light of global change, vectors that are currently not well-documented could become increasingly important, requiring tools to predict their distributions. One way to achieve this could be to leverage data on related species to inform the distribution of a sparsely documented vector based on the assumption that the environmental niches of related species are not independent. Relatedly, there is a natural dependence of the spatial distribution of a disease on the spatial dependence of its vector. Here, we propose to exploit these correlations by fitting a hierarchical model jointly to data on multiple vector species and their associated human diseases to improve distribution models of sparsely documented species. To demonstrate this approach, we evaluated the ability of twelve models – which differed in their pooling of data from multiple vector species and inclusion of disease data – to improve distribution estimates of sparsely documented vectors. We assessed our models on two simulated datasets, which allowed us to generalize our results and examine their mechanisms. We found that when the focal species is sparsely documented, incorporating data on related vector species reduces uncertainty and improves accuracy by reducing overfitting. When data on vector species are already incorporated, disease data only marginally improve model performance. However, when data on other vectors are not available, disease data can improve model accuracy and reduce overfitting and uncertainty. We then assessed the approach on empirical data on ticks and tick-borne diseases in Florida and found that incorporating data on other vector species improved model performance. This study illustrates the value in exploiting correlated data via joint modeling to improve distribution models of data-limited species.

Keywords: data integration, data-limited, joint species distribution modeling, pooling, vector-borne disease

NORDIC SOCIETY OIKOS

www.ecography.org

Introduction

A key component of understanding the risk of vector-borne disease is accurate characterization of the spatial distribution of its vector(s). Species distribution models (SDMs), which estimate a species' distribution from presence–absence or presence-only data are well-established in biogeography and have been successfully applied to study the distribution of well-documented vector species (Barker and MacIsaac 2022, Kopsco et al. 2022). Practical and logistical circumstances influence the degree to which a given vector is documented. Practically, increased effort is likely to be focused on vectors known to transmit pathogens to humans, and we would expect to have larger historical data sets for species that have been longer established within a region. Similarly, some vectors inhabit landscapes that are more common, easier to sample, or promote human contact, thereby increasing the probability that these species are detected.

As global change continues, vectors that are currently not well-documented could become increasingly relevant. For example, as habitat ranges are altered, we may witness increased contact and spillover between species (Wright et al. 2015, Cumbie et al. 2022), allowing current non-vector species to transmit pathogens. Additionally, changing climates may allow for establishment of vector species in areas where they are currently rare or absent (Caminade et al. 2018, Ogden et al. 2020). Further, landscape alteration, caused by human modification or climate change, could result in increased human contact with currently sporadic or removed vector habitats. Therefore, it is essential that we understand the distributions of vectors for which we have limited presence data.

Numerous approaches exist, such as generalized linear models, generalized additive models, boosted regression trees, and MaxEnt models, to estimate species distributions (Peterson et al. 2011). These techniques leverage associations between environmental factors and locations where a species has been documented, necessarily assuming that the training sample represents the environmental conditions within the species' range (Richmond et al. 2010). Therefore, while SDMs produce accurate estimates when developed with large, unbiased training data sets, accuracy declines both as training size decreases (Williams et al. 2009, Aguilar et al. 2016, Boyd et al. 2023) and sample bias increases (Bean et al. 2012), making them less reliable when applied to sparsely documented species.

Joint species distribution models (JSDMs) extend single SDMs by relaxing the assumption of independence between species within an ecological community. JSDMs account for residual co-occurrence patterns between species not explained by model covariates (Pollock et al. 2014). Hierarchical models, such as species archetype models (Hui et al. 2013) and hierarchical models of species communities (Ovaskainen and Soininen 2011), further exploit multispecies information by clustering species based on environmental responses (Hui et al. 2013), or modeling environmental responses as a function of shared species' traits (Zakharova et al. 2019) or genetic units (Ovaskainen and Soininen 2011, Ovaskainen et al. 2017, Escamilla Molgora et al. 2022). These techniques achieve better accuracy by borrowing strength from common species to inform the responses of less common species. The degree of improvement increases as available presence records decrease (Hui et al. 2013).

We propose fitting an extended, hierarchically pooled JSDM to data on multiple vector species within the same taxonomic family and data on the human diseases they transmit to improve distribution estimates for sparsely documented vector species. This approach is based on two assumptions. First, we assume that there is valuable information about the environmental niche of a species that is shared among species at higher taxonomic levels. Genotypic information has been shown to improve distribution models for other species (Banta et al. 2012, Marcer et al. 2016, Zakharova et al. 2019) and a thorough argument for partial pooling within SDMs has been presented elsewhere (Smith et al. 2019). Second, we assume that due to the dependency of a pathogen on the vector that transmits it, there is valuable information about the spatial distribution of the vector contained in the spatial distributions of the diseases it transmits. Hence, through a hierarchical Bayesian framework, we can extend our JSDM to incorporate epidemiological data.

In this paper, we aim to discern whether these techniques improve distribution estimates for sparsely documented vectors. We predicted that joint-hierarchical pooling of vector species and incorporating disease data will increase model accuracy for a sparsely documented focal species due to reduced overfitting compared to the species-independent model. We fitted our models to empirical data on six tick species, within four genera, from the family Ixodidae (*Amblyomma americanum*, *A. maculatum*, *Dermacentor variabilis*, *Ixodes affinis*, *Ixodes scapularis*, *Rhipicephalus sanguineus*) and three human diseases (anaplasmosis, ehrlichiosis, Lyme disease) in Florida. We utilized two simulated data sets – one with all species well-documented and another with the focal species sparsely documented – to test the generalizability of our results and our understanding of the mechanisms driving them. Ticks in Florida serve as a useful case study because multiple species (with differing levels of documentation) coexist. Further, multiple human diseases transmitted by these ticks co-circulate, and disease data are available at a reasonably fine spatial resolution.

Material and methods

Overview

We evaluated the ability of twelve alternative models to improve distribution estimates of sparsely documented vectors (Fig. 1). The first six models differ in their assumptions regarding how information was pooled (Fig. 2) among species. The second six models mirror the first six, but incorporate human disease data. To generalize results and examine their mechanisms, we analyzed the performance of our models fitted to ten simulations for two simulated data sets, one where all species are well-documented and another where the focal species *I. scapularis* is sparsely documented. We assessed

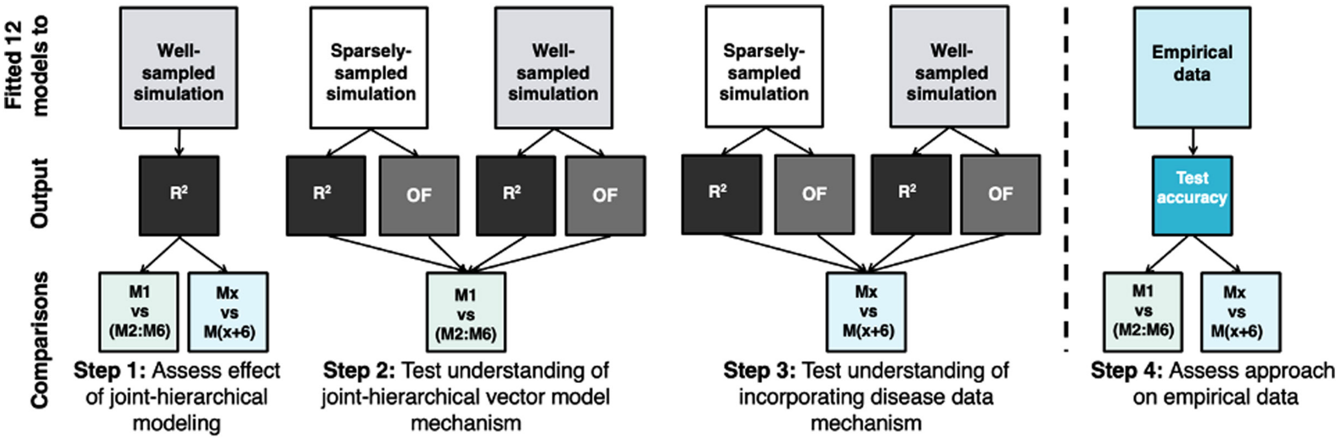

Figure 1.Graphical depiction of the four steps involved in our full analysis. The top row depicts the data to which we fitted our twelve models, the middle row depicts the metrics used to assess model performance, and the bottom row depicts the comparisons used within that step of analysis. R^2 is the coefficient of determination between predicted probability of presence and true probability of presence. OF is a metric of overfitting, and x represents the model number, a value from 1 to 6..

model performance based on each model's ability to recover the distribution from which our focal vector was simulated. We then fitted our models to ten random training data sets derived from empirical presence–absence records of ticks and tick-borne diseases within Florida. We evaluated the average accuracy of model predictions on held-out testing data for a species that we sparsely sampled (*A. maculatum*).

Data

Vector data

Vector presence data were obtained from VectorMap (2020, http://www.vectormap.si.edu) and iNaturalist (https://www.inaturalist.org). Only iNaturalist data considered 'research grade' were included, and we removed duplicates. To obtain absence data, we referenced VectorMap publications and assumed that if a species was not reported at a sampling location, but was included within the study, that the species was absent at that location. To avoid conflating low sampling effort with low vector presence, we based pseudo-absence locations on presence locations from chiggers, fleas, and mites from both databases and the Global Biodiversity Information Facility (www.gbif.org). We used a 1:1 ratio of presence to absence points, which produces the most accurate predicted distribution for regression techniques (Barbet-Massin et al. 2012).

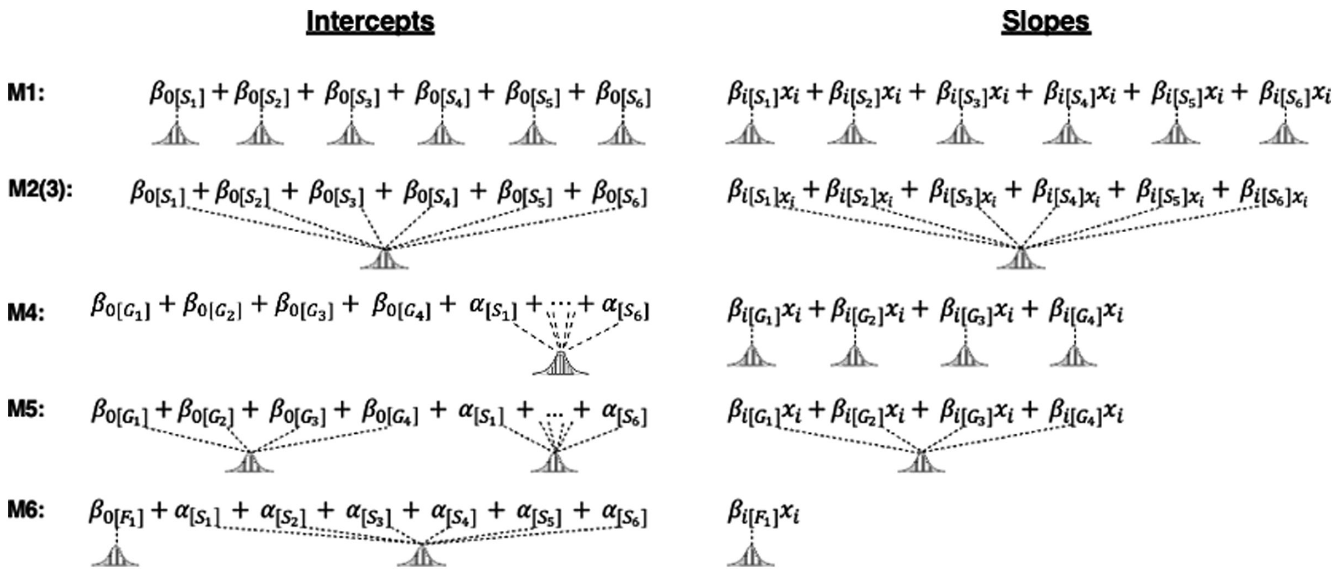

Figure 2. Graphical depiction of how intercepts (left) and response to environmental covariates (right) were modeled for our six distinct vector species models. M1 is the species-independent model, M2 and M3 are the species pooled models, M4 is the genus-independent model, M5 is the genus-pooled model, and M6 is the family-independent model. Per species effects have the additional subscript [S], per genus effects have the additional subscript [G], and per family effects have the additional subscript [F].

We artificially sparsely sampled one species within our empirical data (*A. maculatum*) by including 20% of available presence–absence data in our training set and withholding the rest for testing. The artificial sparse-sampling allowed for a robust testing data set to evaluate model performances. To ensure spatial independence between our training and testing data, data were split using the 'blockCV' package (Valavi et al. 2019) in R ver. 2023.03.0+386 (www.r-project.org). To test the limitations of incorporating disease data, we selected a vector species that does not transmit any of the diseases within our model as our focal species. Empirical sample sizes are given in the Supporting information.

Human disease data

We obtained annual incidence data on three human diseases (anaplasmosis, ehrlichiosis, Lyme disease) from 2011 to 2019 for each county from the Florida Department of Health (https://www.floridahealth.gov/diseases-and-conditions/tick-and-insect-borne-diseases/tick-surveillance.html). We translated this into human disease presence data in a given county in a given year based on whether annual incidence there was greater than zero.

Covariate data

We modeled vector distributions as a function of environmental covariates, which have been linked to tick presence: land cover (Randolph 2000), 30-year average maximum temperature (Ogden et al. 2020), 30-year average precipitation (Ogden et al. 2020), regional Palmer hydrological drought index (Jones and Kitron 2000), normalized differential vegetation index (Randolph 2000), and distance to the nearest waterbody (Kahl and Alidousti 1997). We obtained landcover data from global land cover characteristics database (Loveland et al. 2000), 30-year average climate data from WorldClim (Fick and Hijmans 2017), Palmer hydrological drought index from NOAA (Bushra and Rohli 2017), and normalized difference vegetation index data from USGS Landsat (Vermote et al. 2016). Finally, we obtained waterbody data from the World Wildlife Foundation's global lakes and wetlands database (Lehner and Döll 2004). Pathogen circulation was based on Companion Animal Parasitic Council (https://capcvet.org/maps) data, which report the seroprevalence in canines receiving veterinary treatment. To avoid considering imported cases as indicative of endemicity, we considered a threshold of five annual cases to signal transmission. Finally, to account for under-reporting (Madison-Antenucci et al. 2020), we modeled reporting probability as a function of health insurance coverage and population size. Insurance data were obtained from County Health Rankings (www.countyhealthrankings.org), and population data were obtained from WorldPop (www.worldpop.org).

Simulated data

Our first simulation simulates data for three well-documented species: *A. americanum*, *A. maculatum*, and *D. variabilis*, and a single sparsely documented species: *I. scapularis*. 'Well-documented' is defined as 500 samples and 'sparsely documented' is defined as 30 samples (Supporting information). Our second simulation simulates all four species as well-documented (Supporting information).

Models

Joint-hierarchical vector models

We modeled vector presence for species s at location n_s ($y^T_{sn_s}$) as a Bernoulli function with probability of success $\rho^T_{sn_s}$. The corresponding likelihood function for models fitted to vector data alone is:

$$\ell_V\left(\mathbf{B}_\mathbf{v} \middle| y^T_{sn_s}, \boldsymbol{X}_{\boldsymbol{sVp},\, \boldsymbol{s}\in(1,6)}\right)$$
$$= \prod_{s=1}^{6}\prod_{n_s=1}^{N_s}\left(\rho^T{}_{sn_s}\right)^{y^T_{sn_s}}\left(1-\left(\rho^T{}_{sn_s}\right)\right)^{1-y^T_{sn_s}}. \tag{1}$$

Here, $\mathbf{B}_\mathbf{v}$ is a $c \times s$ matrix with c representing the number of fixed and random effects in our vector presence model and s representing the number of vector species. Our presence–absence data for each species are denoted by $y^T_{sn_s}$. The $N_s \times c$ matrix, $\boldsymbol{X}_{\mathbf{sVp}}$, is composed of 1s (for random-effects) and covariate values (for fixed-effects) at locations of presence–absence data for species s. The number of presence–absence points for species s is denoted by N_s.

We developed six alternative models of vector presence with different pooling structures. We pooled both model intercepts, allowing us to exploit similarities in species presence affected by variables not included in our model, and slopes, allowing us to exploit similarities in species responses to environmental covariates included in our model. Three models (M1–M3) estimate unique parameters for each species and differ based on the assumed covariance between parameters. Two models (M4–M5) estimate unique parameters for each genus and differ based on the assumed covariance between parameters. Finally, one model (M6) estimates a unique parameter for the one family modeled (Fig. 2).

Model 1 (M1) assumes independence of environmental responses for each species. Therefore, statistically, we model presence probability as a linear function of environmental covariates:

$$\operatorname{logit}\left(\rho^T_{sn_s}\right) = \beta_{0[s]} + \sum_{n=1}^{11}\beta_{n[s]}X_{cn_s}, \tag{2}$$

with a unique coefficient $\beta_{n[s]}$ for each environmental covariate n at location x and each species s.

Model 2 (M2) eschews the assumption of independence, instead assuming sufficient niche conservation that we can borrow information on environmental responses across species. Statistically, we estimate a different coefficient for each species s drawn from a shared distribution, such that: $\beta_{n[s]} \sim \operatorname{Normal}\left(\bar{\beta}_n, \sigma_n\right)$ for $c \in (0,11)$, $s \in (0,11)$.

Model 3 (M3) is a phylogenetically informed extension of M2, which explicitly incorporates phylogenetic information and assumes that the similarity in environmental responses between species is inversely proportional to their genetic distance. Genetic distance is defined as the sum of branch lengths between taxa based on a reconstructed hard tick phylogeny (Charrier et al. 2019). Statistically, this is achieved through incorporating a phylogenetic covariance matrix ($\boldsymbol{K}_{nij}$) that informs the random error of each environmental response:

$$\boldsymbol{K}_{nij} = \eta_n^2 \exp\left(-\phi_n^2 \boldsymbol{E}_{ij}\right) + \delta_{ij}\left(0.01\right) \tag{3}$$

$$\delta_{ij} = \begin{cases} 0 \text{ if } i \neq j \\ 1 \text{ if } i = j \end{cases}.$$

Here $\mathbf{E}$ is the phylogenetic distance matrix, η_n^2 represents maximum covariance between two species, and ϕ_n represents rate of decline in similarity.

Model 4 (M4) assumes niche overlap of vectors within the same genus but that the responses among genera are not conserved. Statistically, this is achieved by modeling presence probability as a linear function of environmental covariates:

$$\text{logit}\left(\rho_{sn_s}^T\right) = \beta_{0[g]} + \sum_{n=1}^{11} \beta_{n[g]} X_{cn_s} + \alpha_{[s]}, \tag{4}$$

with a unique coefficient estimated for every environmental covariate for each genus, g. We included a random effect ($\alpha_{[s]} \sim \text{Normal}\left(\bar{\alpha}, \sigma_\alpha\right)$ for $s \in (1,6)$) for each species to allow for species-specific deviations from the shared genus intercept.

The relationship between model 5 (M5) and M4 is analogous to that between M2 and M1. Whereas M2 removes the assumption of independence among species, M5 removes the assumption of independence among genera. Statistically, we estimate a different coefficient for each genus drawn from a shared distribution: such that $\beta_{n[g]} \sim \text{Normal}\left(\bar{\beta}_n, \sigma_n\right)$ for $c \in (0,11)$, $g \in (1,4)$.

Finally, model 6 (M6) assumes all vectors within the same family share environmental responses. Statistically, this is achieved by modeling presence probability as a linear function of environmental covariates:

$$\text{logit}\left(\rho_{sn_s}^T\right) = \beta_{0[f]} + \sum_{n=1}^{11} \beta_{n[f]} X_{cn_s} + \alpha_{[s]}, \tag{5}$$

with a unique coefficient estimated for every environmental covariate for each family. Similar to M4, we included the random effect $\alpha_{[s]}$. M1–M6 were fitted to vector data only.

Incorporating human disease data

Models 7 through 12 (M7–M12) contain identical vector models to M1–M6, respectively. However, M7–M12 are also fitted to human disease data, and require extended structure (Supporting information). Overall, we modeled the presence of human disease d in county i and year j (y_{ijd}) as a Bernoulli process with a probability of success equal to the product of the probability of human disease presence (ψ_{id}) and the probability of the human disease being reported (p_{ijd}), such that:

$$y_{ijd} \sim \text{Bern}\left(\psi_{id}\ p_{ijd}\right). \tag{6}$$

Based on the assumption that the latent probability of disease within each county was constant over our modeled timeframe, we estimated the probability of reporting each year for each county. Consistent with previous work showing an association between health insurance coverage and healthcare-seeking behavior (Ayanian 2000), we modeled the probability of reporting as a linear function of county population size (N_i) and the proportion of individuals in that county with health insurance in year j (S_{ij}), resulting in:

$$\text{logit}\left(p_{ijd}\right) = \beta_{R0[d]} + \beta_{R1[d]} S_{ij} + \beta_{R2[d]} N_i. \tag{7}$$

We modeled the probability that disease d is in county i as a function of the probability that the causative pathogen of disease d is in the county (P_{id}) and the probability that the vector(s) that transmits the pathogen is present in the county (V_{nid}), such that:

$$\begin{aligned} &\text{logit}\left(\psi_{id}\right) \\ &= \beta_{D0[d]} + \beta_{D1[d]} P_{id} + \beta_{D2[d]} V_{1id} + \beta_{D3[d]} V_{2id}. \end{aligned} \tag{8}$$

The probability that the causative pathogen of disease d is in county i was modeled as a linear function of the transformed distance to the closest county in which pathogen d has been reported. Here:

$$\text{logit}\left(P_{id}\right) = \beta_{P0[d]} + \beta_{P1[d]} D'_{id}. \tag{9}$$

The probability of presence was assumed to exponentially decrease as distance increases, according to:

$$D'_{id} = \exp\left(-D_{id}\right). \tag{10}$$

The probability that vector s is in county i (V_{sid}) was calculated as the average probability of vector presence over all grid cells in county i. According to this structure, while disease presence is dependent on vector presence, vector presence is not dependent on disease presence.

The likelihood function for the full model fitted to disease and vector data is given in the Supporting information.

Model fitting

We used a Markov chain Monte Carlo algorithm implemented in RStan (Stan Development Team 2023) to fit our models. For each model–data combination, we ran 4 chains with 40 000 iterations and a burn-in period of 5000. We used normally distributed priors with hyperparameters selected based on prior predictive checks (Supporting information) for all parameters in our linear models. For hyperparameters representing variance and covariance, we used exponential priors to ensure positivity. We inspected trace plots (Supporting information) and utilized the Gelman–Rubin diagnostic to assess convergence with 1.1 as an acceptable threshold. Estimated parameters were evaluated against those used to simulate data for validation (Supporting information).

Model analysis

Simulated data

To test the generalizability of our results, we evaluated the ability of our models to recover the distribution from which our focal species *I. scapularis* was simulated on ten simulated data sets from our sparsely sampled simulation. Model performance was defined as the correlation between model-predicted presence probability and true presence probability (from the simulated distribution) over the model's overall spatial extent (R_o^2). Model uncertainty was defined as the width of 95% Bayesian credibility interval.

We hypothesized that any improvement in performance achieved through joint-hierarchical pooling of vector species or incorporating disease data would be from reducing overfitting compared to the species-independent model. To quantify overfitting, we created an overfitting metric (OF), defined as the percent change between R_T^2 and R_o^2, or:

$$\mathrm{OF} = \frac{R_T^2 - R_O^2}{R_T^2}. \tag{11}$$

We defined R_T^2 as the coefficient of determination between model-predicted presence probability and the presence probability from which the focal species was simulated over the spatial extent of the training data. We predicted that when our focal species was sparsely documented, M1 (the species-independent model) would be more overfitted than our pooled models, because it has the most parameter flexibility. Similarly, we predicted that models fitted to vector data alone would be more overfitted than the corresponding models fitted to vector and human disease data together (i.e. $\mathrm{OF}_{M1} > \mathrm{OF}_{M7}$). In contrast, we predicted that we would not see differences in OF when we trained our models on sufficient data for our focal species. We fitted our models to ten simulated data sets to improve our power to detect small but consistent changes in model performance among alternative models.

Empirical data

We fitted our twelve models to ten training data sets derived from empirical data on six tick species and three tick-borne diseases in Florida. Then, we evaluated the average accuracy of model predictions compared to testing data for our sparsely documented, focal vector species *A. maculatum*. Accuracy is defined as the proportion of true positives (TP) and true negatives (TN) among all possible outcomes (i.e. TP and TN, false positives [FP] and false negatives [FN]), which is calculated as:

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}. \tag{12}$$

We focus our assessment on accuracy because it is intuitive and incorporates presences and absences. To assess the robustness of our conclusions to metric choice, we also calculated area under the curve, kappa statistic, and true skill statistic. We found that relative model performance was consistent across metrics (Supporting information). If hierarchical pooling of data on vector species improved the estimated distributions of our sparsely documented species, we would expect, on average, predictions from the pooled models to be more accurate than predictions from the species-independent model. If incorporating disease data improved the estimated distributions of our sparsely documented species, we would expect predictions from the models fitted to both vector and disease data to be more accurate than predictions from the corresponding models fitted to vector data alone.

Results

Analysis of simulated data

Joint-hierarchical model

To test the effect of pooling vector species when the focal species was sparsely documented, we compared model performance of our pooled models (M2–M6) to model performance of our independent model (M1) over 10 simulation runs. Within the simulation where our focal species was sparsely documented, on average, our pooled models outperformed our species-independent model ($R_o^2 = 0.68$ versus $R_o^2 = 0.41$) and decreased model uncertainty (Uncertainty = 0.41 versus Uncertainty = 0.60). Importantly, within this simulation, M4 (the genus-independent model) is a species-independent model because our focal species *I. scapularis* is the only species within its genus in our simulated data. With a single species within the genus, pooling at the genus level is equivalent to pooling at the species level. This constraint was due to insufficient presence records to estimate the distribution of the second species within the genus, and is a limitation of our study. When we classify M4 as a species-independent model, the improvement derived from hierarchal pooling of vector species increases ($R_O^2 = 0.77$ versus $R_O^2 = 0.38$). The improved performance derived from pooling vector data when the focal species is sparsely documented is consistent with our predictions. Table 1 and Fig. 3 summarize the performances of M1–M6 fitted to simulated data. Differences between predicted distributions among models are depicted in Fig. 4.

Table 1. Average R_O^2 (and range) between model-predicted distribution and true distribution, average overfitting (OF) (and range), and uncertainty for our focal species *Ixodes scapularis* over 10 runs of the sparsely sampled simulation. R^2 values measure agreement between the 'true' distribution used within our simulation and the distribution predicted by our model, with higher R^2 values indicating stronger agreement. OF values measure the percent change in R^2 between the geographic areas represented in the training data and the full geographic range of our predictions. Higher OF values indicate that the model is overfitted, whereas OF values at or below zero indicate that the model is not overfitted to the training data.

Model	Structure	Data sources	Simulation	Avg R^2	Avg OF	Uncertainty
M1	species independent	vector	sparsely sampled	0.41 (0–0.68)	0.29 (−0.32 to 0.99)	0.60
M2	species pooled	vector	sparsely sampled	0.77 (0.53–0.92)	−0.06 (−0.56 to 0.16)	0.46
M3	species pooed	vector	sparsely sampled	0.78 (0.54–0.95)	0.01 (−0.24 to 0.18)	0.42
M4	genus independent	vector	sparsely sampled	0.34 (0–0.57)	0.38 (−0.35 to 1.13)	0.60
M5	genus pooled	vector	sparsely sampled	0.76 (0.54–0.92)	0 (−0.31 to 0.18)	0.47
M6	family independent	vector	sparsely sampled	0.77 (0.55–0.95)	−0.05 (−0.53 to 0.16)	0.23
M7	species independent	vector, disease	sparsely sampled	0.52 (0.05–0.88)	0.23 (0–0.82)	0.58
M8	species pooled	vector, disease	sparsely sampled	0.80 (0.48–0.94)	−0.06 (−0.47 to 0.16)	0.45
M9	species pooed	vector, disease	sparsely sampled	0.80 (0.52–0.95)	0 (−0.23 to 0.17)	0.40
M10	genus independent	vector, disease	sparsely sampled	0.51 (0–0.91)	0.25 (0 to 1.09)	0.59
M11	genus pooled	vector, disease	sparsely sampled	0.77 (0.44–0.95)	−0.03 (−0.48 to 0.17)	0.43
M12	family independent	vector, disease	sparsely sampled	0.80 (0.55–0.91)	−0.04 (−0.54 to 0.16)	0.29

To determine the specific reasons that some models performed better than others, we quantified overfitting. On average, our species-independent model ($OF_i = 0.29$) was more overfitted than our pooled models ($OF_p = 0.06$). When we classified M4 as a species-independent model, the overfitting increased in our independent models ($OF_i = 0.34$) and decreased in our pooled models ($OF_p = -0.03$). An overfitting value of 0.34 indicates that our model predictions were 34% more accurate within the spatial extent of the training data than over the full spatial extent of our model, while an overfitting value of −0.03 indicates our model predictions were 3% more accurate over the model's full spatial extent, meaning the model was not overfitted to the training data. These results are consistent with our prediction that a reduction in overfitting from pooling data is driving differences in model performances (Table 1).

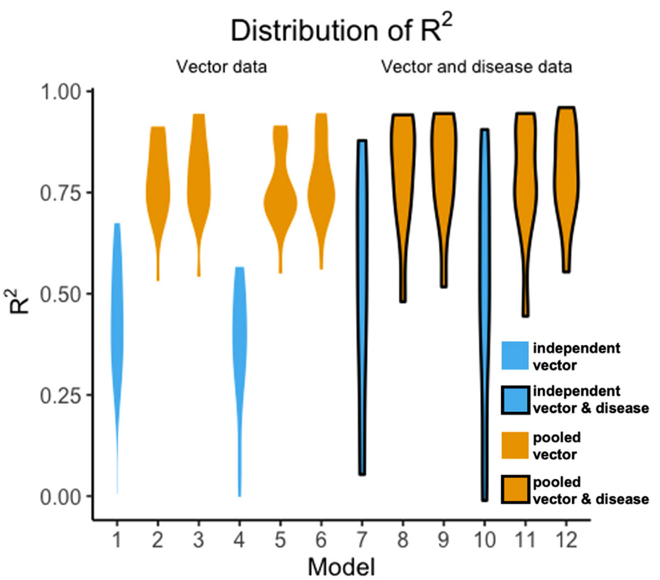

Figure 3. Distribution of R_O^2 between predicted and true probability distributions for our sparsely documented species, *Ixodes scapularis*, for all 12 models over 10 simulation runs. Models with the black outline are fitted to vector data and human disease data. Models to the left of the dashed line with no outline are fitted to vector data alone.

To isolate the effect of sample size, as a control, we fitted the same six models to a simulation where our focal species was well-documented. Within this simulation, there was no difference in performance ($R_{Oi}^2 = 0.87$ versus $R_{Op}^2 = 0.88$), and a small decrease in uncertainty ($R_{Oi}^2 = 0.22$ versus $R_{Op}^2 = 0.19$) between our species-independent model and our pooled models (Supporting information). Similarly, the difference in overfitting between the independent and pooled models was minor (Supporting information). This result is consistent with our hypothesis and reflects that models fitted to sufficient data can accurately estimate the distribution of the species, not just the distribution in the training data.

To determine if there was an optimal pooling structure, we compared performance across our six models. Within both of our simulations, M3 (the phylogenetically informed model) performed best. However, the improvement in M3 over other pooled models was marginal: $R_O^2 = 0.84$ versus $R_O^2 = 0.82$) (Table 1, Fig. 3, Supporting information).

Incorporating human disease data

To test the effect of incorporating human disease data, we compared the average R_O^2 of each version of the model with and without human disease data (e.g. M1 versus M7) over 10 simulation runs. Within the simulation where our focal species was sparsely documented, incorporating human disease data increased R_O^2 for the estimated focal-vector distribution by 0.06. To determine the utility of incorporating disease data when additional vector data are not available, we compared the difference in model performance derived from adding disease data in our pooled (M2, M3, and M6) and species-independent (M1 and M4) models. In our pooled models, incorporating disease data generated an average improvement of 0.02. In our species-independent models, incorporating disease data generated an average improvement of 0.14

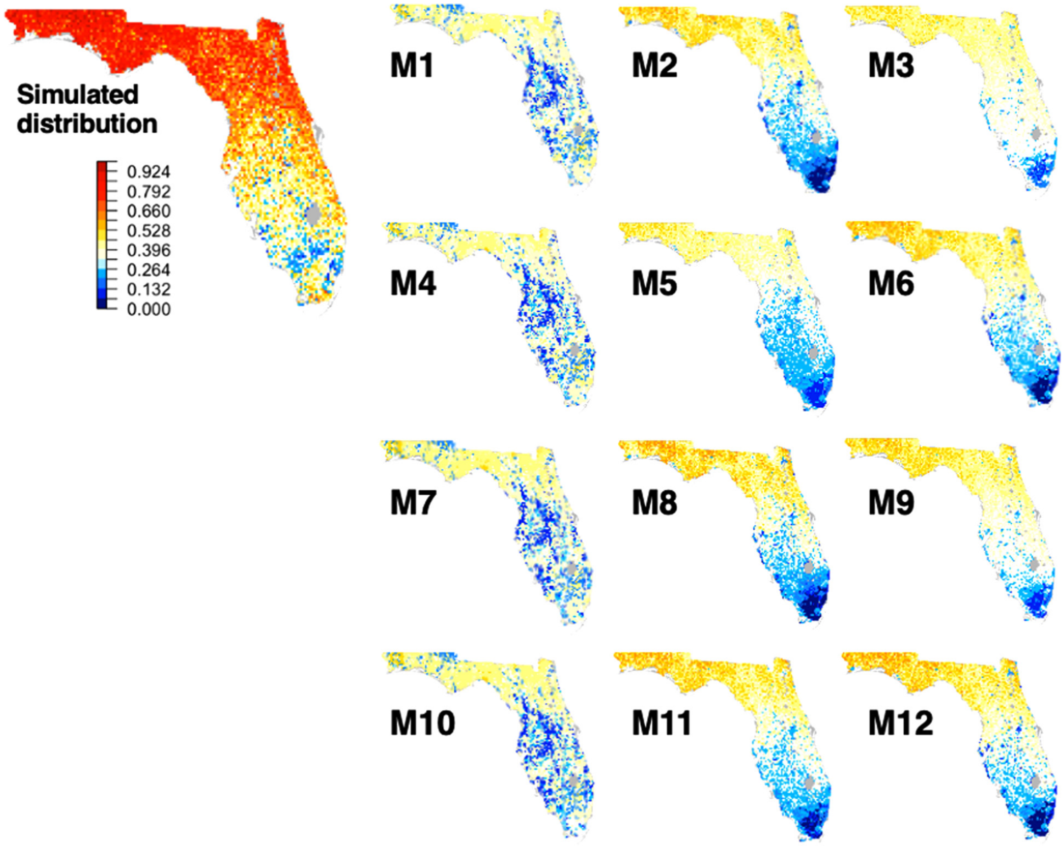

Figure 4. True vector distribution within our simulation (left) and predicted distributions (right) of *Ixodes scapularis* when sparsely documented for each of our twelve models from a randomly selected simulation run. Outputs in the first column are from species-independent models. Outputs in the second and third columns are from pooled models.

(Table 1, Fig. 3). This result demonstrates that incorporating disease data improves performance more when data on additional vector species are not available. Across all models, incorporating disease data marginally reduced uncertainty. Differences in predicted distributions among models are depicted in Fig. 4.

To determine the specific reasons for this improvement, we calculated an overfitting metric for models fitted to both vector and human disease data. Within the species-independent models, on average, overfitting was 0.24, demonstrating that adding disease data reduced overfitting by 0.10 (OF = 0.34 without disease data versus OF = 0.24 with disease data). Within our pooled models, overfitting was equivalent with and without disease data (OF = 0.03). This result demonstrates that incorporating disease data leads to model improvement by reducing overfitting within our species-independent models (Table 1).

As a control, we made the same model comparisons in a simulation where our focal species was well-documented. For our focal species, there was on average a 0.01 improvement in model performance (e.g. R^2_O of M7 versus R^2_O of M1) in the well-sampled simulation when adding disease data and no change in model uncertainty. Average overfitting over all models was 0.01. Similarly, when all species were well-documented, there was no change in overfitting between the models with and without disease data, demonstrating that disease data can benefit distribution models for a sparsely documented species, but only marginally affects model predictions for well-documented species (Supporting information).

Analysis of empirical data

To determine the utility of pooling data to improve the predicted distribution of our sparsely documented species *A. maculatum*, we compared the accuracy of the species-independent models to the average accuracy of the pooled models over ten test-train splits. The average accuracy of the pooled models ($Acc_p = 0.59$) was higher than the average accuracy of the species-independent model ($Acc_i = 0.53$), and the average uncertainty of the pooled models ($Uncertainty_p = 0.28$) was lower than the uncertainty of the independent models ($Uncertainty_i = 0.65$). This result is consistent with our simulated results and predictions. To assess the impact of incorporating disease data to improve the predicted distribution of our sparsely documented species, we compared model accuracy with and without disease data across ten test-train splits. We found no effect of incorporating disease data on accuracy or uncertainty, likely due to *A. maculatum* not transmitting the diseases in the model (Table 2).

Discussion

In this study, we evaluated the utility of exploiting ecological correlations between related vector species and the human diseases they transmit to improve spatial distribution models of sparsely documented vector species. Due to the assumed spatial correlation between the distributions of sparsely documented vector species, more well-sampled vector species, and vector-borne human disease, we predicted that models that jointly considered all of these data would reduce overfitting and improve performance. We found that incorporating data on additional vector species improved model performance for sparsely documented vectors. When the sparsely documented species did not transmit the diseases in the model, incorporating human disease data had no effect on model performance. When the sparsely documented vector did transmit the diseases in the model, incorporating disease data alone improved model performance by reducing overfitting, particularly when data on other vector species were not incorporated.

When fitted to simulated data and empirical data, our pooled distribution models for a sparsely documented species outperformed our species-independent models. The average overfitting (OF) value of our species-independent models confirms that, unlike our pooled models, the species-independent models are overfitted and therefore less able to reliably make out-of-context predictions. Past studies have found niche overlap both within a tick genus (Estrada Peña 2019) and across tick families (Peralbo-Moreno et al. 2022), and niche divergence across tick genera (Tkadlec et al. 2018). While we expected that species within the same genus may have conserved niches, we did not expect to see as much similarity between species related only at the family level. This result could suggest trait conservation at the family level for the specific species within our model.

While our findings on the benefits of including data on additional vector species within a taxonomic family were consistent across our empirical and simulated analyses, we found a larger effect of pooling within our simulations. A plausible explanation is that the estimated independent distribution of *I. scapularis* was more similar to the estimated independent distributions of the other vectors represented in our data than that of *A. maculatum* (Supporting information). This underscores an important limitation of our approach: the effectiveness of pooling among vector species depends on similarity of their individual distributions.

In our simulations in which our sparsely documented species *I. scapularis* was a vector of the diseases in the model, disease data marginally improved model performance. However, there was a substantial difference in improvement associated with disease data between our species-independent and pooled models. Specifically, our species-independent models saw a larger reduction in overfitting than our pooled models when fitted to disease data. This result suggests that when the sparsely documented species is a vector of the diseases in the model, and data on other related vector species are not available, disease data can improve model performance by decreasing overfitting. Disease data are only marginally informative, however, if presence data on related vector species are already incorporated (see Supporting information for detail of how incorporating vector versus disease data differently constrains parameters). It is plausible that the limited utility of disease data, even when informing the distribution of the vector that transmits the disease in the model, arises because disease data are only indirectly related to vector data. For example, although disease presence is affected by vector presence, it also requires the presence of the causative pathogen, human interaction with the vector, and diagnosis of the disease. Additionally, since one disease (ehrlichiosis) is caused by both our focal vector *I. scapularis* and a second vector, there may be only a weak correlation between disease and focal-vector presence. Disease data had little to no effect on model performance when the focal species was well-documented and the model was trained on adequate data, or when the focal, sparsely documented vector species did not transmit the diseases included in the model. This reinforces the view that incorporating disease data may be most beneficial for distribution models of sparsely documented species that lack well-documented, related species to help inform their distribution, and transmit well-documented diseases. However, even when that focal species was well-sampled, or did not transmit the diseases in the model, disease data did not introduce additional bias or reduce model performance.

Table 2. Average accuracy (and range) and uncertainty for our focal species *Amblyomma maculatum* for all models over 10 test-train splits of empirical data.

Model	Structure	Data sources	Accuracy (range)	Uncertainty
M1	species independent	vector	0.53 (0.50–0.60)	0.65
M2	species pooled	vector	0.58 (0.53–0.62)	0.51
M3	species pooed	vector	0.59 (0.5–0.61)	0.28
M4	genus independent	vector	0.58 (0.55–0.60)	0.20
M5	genus pooled	vector	0.58 (0.55–0.60)	0.20
M6	family independent	vector	0.60 (0.58–0.62)	0.20
M7	species independent	vector, disease	0.53 (0.51–0.60)	0.65
M8	species pooled	vector, disease	0.58 (0.54–0.62)	0.51
M9	species pooed	vector, disease	0.59 (0.57–0.61)	0.28
M10	genus independent	vector, disease	0.58 (0.55–0.60)	0.20
M11	genus pooled	vector, disease	0.58 (0.55–0.60)	0.20
M12	family independent	vector, disease	0.60 (0.58–0.62)	0.21

Within our simulations, we only simulated a single species within the *Ixodes* genus, causing our genus-independent model to reduce to a species-independent model, which was a limitation of our study. We did so to prevent simulating species from biased distributions caused by insufficient presence records in our empirical training data. This was particularly important since we were specifically evaluating taxonomic correlation in each species' environmental response, and any correlation could be diluted by simulating from a biased estimated distribution.

Overall, we have shown the benefits of joint modeling of additional vector and human disease data towards improving distribution estimates for a sparsely documented vector species. Others have proposed hierarchical pooling on theoretical grounds (Smith et al. 2019). Our study adds value by evaluating when this technique is most useful (i.e. when a vector is sparsely documented) and comparing models pooled at different taxonomic levels. Similar to other works supplementing SDMs with data on other species (Fithian et al. 2014, Valle and Tucker Lima 2014) or genetic information (Marcer et al. 2016, Godoy et al. 2018, Wang et al. 2019), our results highlight the potential of leveraging data from related species. To the best of our knowledge, this is the first study to utilize spatially correlated epidemiological data to inform an SDM of a disease vector. Incorporating data on other vector species within the family proved to be most useful. Incorporating disease data improved model performance and reduced model uncertainty if the sparsely documented species transmitted the diseases in the model, particularly when data on other species were not available. Further, even when the sparsely documented species did not transmit the diseases in the model, incorporating disease data did not hurt model performance. Therefore, epidemiological data could provide an untapped, supplemental data source towards informing distribution models of data-limited vector or obligate host/reservoir species.

Acknowledgements – We thank Margaret Elliott and Brooke Rodriguez for assisting with human disease data collection. This manuscript was improved by the thoughtful comments of two anonymous reviewers.

Funding – This work was supported by the US Army Medical Research and Development Command under contract no. W81XWH-21-C-0001, no. W81XWH-22-C-0093, and no. HT9425-23-C-0059. The views and opinions and/or findings contained in the report are those of the authors and should not be construed as an official Department of the Army of Navy position, policy or decisions unless so designated by other documentation. TAP also received support from the NIH National Institute of General Medical Sciences R35 MIRA program (grant no. R35GM143029).

Author contributions

Stacy Mowry: Conceptualization (lead); Data curation (lead); Formal analysis (lead); Methodology (lead); Writing – original draft (lead). **Sean Moore**: Supervision (supporting); Writing – review and editing (supporting). **Nicole L. Achee**: Conceptualization (supporting); Funding acquisition (equal). **Benedicte Fustec**: Data curation (supporting); Writing – review and editing (supporting). **T. Alex Perkins**: Conceptualization (supporting); Formal analysis (supporting); Funding acquisition (equal); Methodology (supporting); Supervision (lead); Writing – review and editing (lead).

Transparent peer review

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/ecog.07253.

Data availability statement

Data are available from Zenodo: https://doi.org/10.5281/zenodo.10974724 (Mowry et al. 2024).

Supporting information

The Supporting information associated with this article is available with the online version.

References

Aguilar, G., Waqa-Sakiti, H. and Winder, L. 2016. Using predicted locations and an ensemble approach to address sparse data sets for species distribution modelling: long-horned beetles (Cerambycidae) of the Fiji Islands. – Unitec ePress.

Ayanian, J. Z. 2000. Unmet health needs of uninsured adults in the United States. – JAMA 284: 2061.

Banta, J. A., Ehrenreich, I. M., Gerard, S., Chou, L., Wilczek, A., Schmitt, J., Kover, P. X. and Purugganan, M. D. 2012. Climate envelope modelling reveals intraspecific relationships among flowering phenology, niche breadth and potential range size in *Arabidopsis thaliana*. – Ecol. Lett. 15: 769–777.

Barbet-Massin, M., Jiguet, F., Albert, C. and Thuiller, W. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? – Methods Ecol. Evol. 3: 327–338.

Barker, J. R. and MacIsaac, H. J. 2022. Species distribution models applied to mosquitoes: use, quality assessment, and recommendations for best practice. – Ecol. Modell. 472: 110073.

Bean, W. T., Stafford, R. and Brashares, J. S. 2012. The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. – Ecography 35: 250–258.

Boyd, R. J., Harvey, M., Roy, D. B., Barber, T., Haysom, K. A., Macadam, C. R., Morris, R. K. A., Palmer, C., Palmer, S., Preston, C. D., Taylor, P., Ward, R., Ball, S. G. and Pescott, O. L. 2023. Causal inference and large-scale expert validation shed light on the drivers of SDM accuracy and variance. – Divers. Distrib. 29: 774–784.

Bushra, N. and Rohli, R. V. 2017. County-level drought indices The Palmer Drought Severity Index (PDSI) and Palmer Hydrological Drought Index (PHDI). – US Geological Survey.

Caminade, C., McIntyre, K. M. and Jones, A. E. 2018. Impact of recent and future climate change on vector-borne diseases. – Ann. N. Y. Acad. Sci. 1436: 157–173.

Charrier, N. P., Hermouet, A., Hervet, C. 2019. A transcriptome-based phylogenetic study of hard ticks (Ixodidae). – Sci. Rep. 9: 12923.

Cumbie, A. N., Trimble, R. N. and Eastwood, G. 2022. Pathogen spillover to an invasive tick species: first detection of bourbon virus in *Haemaphysalis longicornis* in the United States. – Pathogens 11: 454.

Escamilla Molgora, J. M., Seda, L., Diggle, P. J. and Atkinson, P. M. 2022. A taxonomic-based joint species distribution model for presence-only data. – J. R. Soc. Interface 19: 20210681.

Estrada Peña, A. 2019. Review of 'species delimitation of the dermacentor ticks based on phylogenetic clustering and niche modeling.

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

Fithian, W., Elith, J., Hastie, T. and Keith, D. A. 2014. Bias correction in species distribution models: pooling survey and collection data for multiple species. – Methods Ecol. Evol. 6: 424–438.

Godoy, B. S., Camargos, L. M. and Lodi, S. 2018. When phylogeny and ecology meet: modeling the occurrence of Trichoptera with environmental and phylogenetic data. – Ecol. Evol. 8: 5313–5322.

Hui, F., Warton, D., Foster, S. and Dunstan, P. 2013. To mix or not to mix: comparing the predictive performance of mixture models vs. separate species distribution models. – Ecology 94: 1913–1919.

Jones, C. J. and Kitron, U. D. 2000. Populations of *Ixodes scapularis* (Acari: Ixodidae) are modulated by drought at a lyme disease focus in Illinois. – J. Med. Entomol. 37: 408–415.

Kahl, O. and Alidousti, I. 1997. Bodies of liquid water as a source of water gain for Ixodes ricinus ticks (Acari: Ixodidae). – Exp. Appl. Acarol. 21: 731–746.

Kopsco, H. L., Smith, R. L. and Halsey, S. J. 2022. A scoping review of species distribution modeling methods for tick vectors. – Front. Ecol. Evol. 10: 893016.

Lehner, B. and Döll, P. 2004. Development and validation of a global database of lakes, reservoirs and wetlands. – J. Hydrol. 296: 1–22.

Loveland, T. R., Reed, B. C., Brown, J. F., Ohlen, D. O., Zhu, J., Yang, L. and Merchant, J. W. 2000. Development of a global land cover characteristics database and IGBP DISCover from 1-km AVHRR data. – Int. J. Remote Sens. 21: 1303–1330.

Madison-Antenucci, S., Kramer, L. D., Gebhardt, L. L. and Kauffman, E. 2020. Emerging tick-borne diseases. – Clin. Microbiol. Rev. 33: e00083-18.

Marcer, A., Mendez-Vigo, B., Alonso-Blanco, C. and Pico, F. X. 2016. Tackling intraspecific genetic structure in distribution models better reflects species geographical range. – Ecol. Evol. 6: 2084–2097.

Mowry, S., Moore, S., Achee, N. L., Fustec, B. and Alex Perkins, T. 2024. Data from: Improving distribution models of sparsely documented disease vectors by incorporating information on related species via joint modeling. – Zenodo Digital Repository, https://doi.org/10.5281/zenodo.10974724.

Ogden, N. H., Ben Beard, C., Ginsberg, H. S. and Tsao, J. I. 2020. Possible effects of climate change on ixodid ticks and the pathogens they transmit: predictions and observations. – J. Med. Entomol. 58: 1536–1545.

Ovaskainen, O. and Soininen, J. 2011. Making more out of sparse data: hierarchical modeling of species communities. – Ecology 92: 289–295.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume B., Duan, L., Dunson, D., Roslin, T. and Abrego, N. 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. – Ecol. Lett. 20: 561–576.

Peralbo-Moreno, A., Baz-Flores, S., Cuadrado-Matías, R., Barroso, P., Triguero-Ocaña, R., Jiménez-Ruiz, S., Herraiz, C., Ruiz-Rodríguez, C., Acevedo, P. and Ruiz-Fons, F. 2022. Environmental factors driving fine-scale ixodid tick abundance patterns. – Sci. Total Environ. 853: 158633.

Peterson, A. T., Soberon, J., Peasrson, R. G., Anderson, R. P., Martinez-Meyer, E., Nakamura, M. and Araujo, M. B. 2011. Ecological niches and geographic distributions (MPB-49). – Princeton Univ. Press.

Pollock, L. J., Tingley, R., Morris, W., Golding, N., O'Hara, R. B., Parris, K., Vesk, P. A. and McCarthy, M. A. 2014. Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (JSDM). – Methods Ecol. Evol. 5: 397–406.

Randolph, S. E. 2000. Ticks and tick-borne disease systems in space and from space. – Adv. Parasitol. 47: 217–243.

Richmond, O. M., McEntee, J. P., Hijmans, R. J. and Brashares, J. S. 2010. Is the climate right for Pleistocene rewilding? using species distribution models to extrapolate climatic suitability for mammals across continents. – PLoS One 5: e12899.

Smith, A. B., Godsoe, W., Rodriguez-Sanchez, F., Wang, H. and Warren, D. 2019. Niche estimation above and below the species level. – Trends Ecol. Evol. 34: 260–273.

Stan Development Team 2023. RStan: the R interface to Stan. – R package ver. 2.21.8, https://mc-stan.org.

Tkadlec, E., Vaclavik, T., Kubelova, M. and Siroky, P. 2018. Negative spatial covariation in abundance of two European ticks: diverging niche preferences or biotic interaction? – Ecol. Entomol. 43: 804–812.

Valavi, R., Elith, J., Lahoz-Monfort, J. J. and Guillera-Arroita, G. 2019. blockCV: an R package for generating spatially or environmentally separated folds for *k*-fold cross-validation of species distribution models. – Methods Ecol. Evol. 10: 225–232.

Valle, D. and Tucker Lima, J. M. 2014. Large-scale drivers of malaria and priority areas for prevention and control in the Brazilian amazon region using a novel multi-pathogen geospatial model. – Malar. J. 13: 443.

VectorMap Data Portal 2020. National museum of natural history, Smithsonian Institution. – Univ. of Queensland Insect Collection, Museum of Vertebrate Zoology, Univ. of California, http://www.vectormap.si.edu.

Vermote, E., Justice, C., Claverie, M. and Franch, B. 2016. Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. – Remote Sens. Environ. 185: 46–56.

Wang, F., Wang, D., Guo, G., Hu, Y., Wei, J. and Liu, J. 2019. Species delimitation of the Dermacentor ticks based on phylogenetic clustering and niche modeling. – PeerJ. 7: e6911.

Williams, J. N., Seo, C., Thorne, J., Nelson, J. K., Erwin, S., O'Brien, J. M. and Schwartz, M. W. 2009. Using species distribution models to predict new occurrences for rare plants. – Divers. Distrib. 15: 565–576.

Wright, C. L., Sonenshine, D. E., Gaff, H. D. and Hynes, W. L. 2015. *Rickettsia parkeri* transmission to *Amblyomma americanum* by cofeeding with *Amblyomma maculatum* (Acari: Ixodidae) and potential for spillover. – J. Med. Entomol. 52: 1090–1095.

Zakharova, L., Meyer, K. M. and Seifan, M. 2019. Trait-based modelling in ecology: a review of two decades of research. – Ecol. Modell. 407: 108703.