

Appendix A

In reference to “Estimating the density of questing *Ixodes scapularis* nymphs in the eastern United States using climate and land cover data” in *Ticks and Tick-Borne Diseases*

Karen M Holcomb, Erik Foster, Rebecca J Eisen

This Appendix contains Text A.1, three tables (Tables A.1-A.3), and ten figures (Figure A.1-A.10). The following provides a brief description of each.

- Text A.1 contains additional mathematical details on the zero-inflated negative binomial (ZINB) model.
- Table A.1 provides further details on correlations within and between groups of variables obtained with hierarchical clustering.
- Table A.2 provides a state-level summary of collections of questing *I. scapularis* nymphs reported to the ArboNET Tick Module (2004-2023) without distance or time-of-year exclusion criteria applied.
- Figure A.1 illustrates spatial distribution of sampling intensity metrics by included counties.
- Figure A.2 provides a visual representation of the candidate covariates for the ZINB model and Table A.3 provides a numerical comparison of the mean and range of these covariates in counties included in model fitting and across counties in the eastern United States.
- Figure A.3 illustrates counties as assigned as a structural and random zero by the fitted ZINB model.
- Figure A.4 illustrates the univariate and multivariate environmental similarity surface (MESS) between covariates in counties used to fit the model and those in the rest of the eastern US.
- Figure A.5 provides a comparison of observed and predicted category of collection probability using optimal threshold identified with ROC curve analysis as well as a contingency table summarizing the observed and predicted categories.
- Figure A.6 provides a visual comparison of the predicted probability questing nymphs would be collected and observed presence/absence of collected nymphs using the full dataset in the ArboNET Tick module (i.e., not applying the exclusion criteria). The figure also provides a contingency table comparing the predicted and observed categorization of collection probability.
- Figure A.7 illustrates the relationship between standardized χ^2 Pearson residuals from the final ZINB model and sampling intensity metrics.
- Figure A.8 provides a map of estimated DON dichotomized at one questing nymph collected per 1,000 m² and considering the 95% CI of estimated DON.
- Figure A.9 illustrates the relationship between predicted and observed relative risk categorization and sampling intensity metrics.
- Figure A.10 illustrates standardized χ^2 Pearson residuals relative to observed and predicted relative risk categories.

Text A.1. Mathematical details

A zero-inflated negative binomial (ZINB) model for predicting the density of questing *Ixodes scapularis* nymphs collected ($\hat{y}$) can be thought of as a combination of a logistic regression model predicting the probability questing *I. scapularis* nymphs are absent ($\hat{p}$) and a negative binomial regression model predicting the density of nymphs collected ($\hat{\mu}$):

$$\hat{y} = (\hat{p})(0) + (1 - \hat{p})(\hat{\mu}). \quad (\text{A.1})$$

Each piece of the ZINB model uses a set of covariates for predicting the respective values such that the logistic piece has the form:

$$\hat{p} = \frac{\exp(\hat{\gamma}_0 + \hat{\gamma}_1 z_1 + \hat{\gamma}_2 z_2 + \dots)}{1 + \exp(\hat{\gamma}_0 + \hat{\gamma}_1 z_1 + \hat{\gamma}_2 z_2 + \dots)} \quad (\text{A.2})$$

and the negative binomial piece has the form:

$$\hat{\mu} = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots) \quad (\text{A.3})$$

where the $\hat{\beta}$'s and $\hat{\gamma}$'s are estimated coefficients for the respective negative binomial and zero-inflated components of the model and x 's and z 's are predictors for the respective negative binomial and zero-inflated pieces of the model. See Table 2 for estimated coefficients in the final ZINB model.

Table A.1. Clusters and average correlations of variables resulting from hierarchical clustering using 1-|correlation| as the distance measure between pairs of variables. Hierarchical clustering used for covariate selection to obtain a subset of rather uncorrelated variables for model fitting. See Section 2.2

Environmental Covariates for further details.

Cluster	Variable	Average within-cluster correlation	Average correlation to closest cluster	Closest cluster
1	Precipitation of driest quarter (Bio17)	0.82	0.63	3
	Precipitation of driest month (Bio14)	0.82	0.62	3
	Precipitation of coldest quarter (Bio19)	0.80	0.69	3
	Precipitation seasonality (Bio15)	0.75	0.46	3
	Annual precipitation (Bio12)	0.75	0.73	3
	Percent forest	0.60	0.36	5
	Percent crops	0.55	0.32	5
2	Percent open space & low intensity urban	0.75	0.22	1
	Moderate/high intensity urban	0.75	0.17	4
3	Mean temperature of coldest quarter (Bio11)	0.80	0.54	1
	Min temperature of warmest month (Bio6)	0.78	0.55	1
	Temperature seasonality (Bio4)	0.76	0.71	1
	Annual mean temperature (Bio1)	0.75	0.62	4
	Isothermality (Bio3)	0.74	0.59	1
	Mean temperature of driest quarter (Bio9)	0.73	0.57	1
	Temperature annual range (Bio7)	0.72	0.69	1
	Precipitation of wettest quarter (Bio16)	0.67	0.44	1
	Precipitation of wettest month (Bio13)	0.67	0.42	1
	Precipitation of warmest quarter (Bio18)	0.50	0.33	1
4	Max temperature of warmest month (Bio5)	0.61	0.55	3
	Mean temperature of warmest quarter (Bio10)	0.58	0.60	3
	Mean diurnal range (Bio2)	0.38	0.29	3
	Mean elevation	0.36	0.22	2
5	Mean temperature of wettest quarter (Bio8)	1.00	0.33	1

Table A.2. Summary of reported collections of questing *I. scapularis* nymphs to the ArboNET Tick Module (2004-2023). Collections used either dragging or flagging methods.

Region ⁺	State	No. (%) counties sampled	Total no. nymphs collected	Total no. collection events
Great Plains	Kansas	13 (12.4)	1	519
	North Dakota	5 (9.4)	0	619
	Nebraska	28 (30.1)	43	149
	Oklahoma	17 (22.1)	5	267
	South Dakota	9 (13.6)	14	60
	Texas	8 (3.2)	0	254
Northeast	Connecticut	8 (100)	4,901	1,205
	District of Columbia	1 (100)	722	23
	Massachusetts	1 (7.1)	846	301
	Maryland	8 (33.3)	1,901	374
	Maine	16 (100)	2,646	661
	New Hampshire	10 (100)	293	471
	New Jersey	21 (100)	6,605	752
	New York	55 (88.7)	26,640	1,702
	Pennsylvania	67 (100)	13,728	4,302
	Rhode Island	2 (40.0)	108	21
	Virginia	41 (30.8)	6,830	761
	Vermont	12 (85.7)	1,828	1,179
	West Virginia	40 (72.7)	1,265	544
Southeast	Alabama	11 (16.4)	25	606
	Arkansas	2 (2.7)	0	6
	Florida	18 (26.9)	38	499
	Georgia	18 (11.3)	1	249
	Louisiana	1 (1.6)	0	1
	Mississippi	55 (67.1)	1	230
	North Carolina	37 (37.0)	1,381	289
	South Carolina	18 (39.1)	7	63
	Tennessee	33 (34.7)	260	546
Upper Midwest	Iowa	42 (42.4)	225	261
	Illinois	64 (62.7)	682	494
	Indiana	91 (98.9)	2,451	741
	Kentucky	13 (10.8)	8	42
	Michigan	73 (88.0)	2,999	1,258
	Minnesota	69 (79.3)	7,318	800
	Missouri	45 (39.1)	7	122
	Ohio	81 (92.0)	501	539
	Wisconsin	42 (58.3)	7,085	831

⁺See Fig. 1A for map of regions in the eastern US.

Table A.3. Range of covariates considered for describing and predicting the peak density of questing *Ixodes scapularis* nymphs. Ranges shown for the 668 counties with observations for training the model and the full 2,694 counties across the eastern United States[†] for which we produced estimates.

Covariate	Mean (range) in included counties	Mean (range) across eastern US[†]
Mean diurnal temperature range (Bio2; °C)	11.80 (9.07-14.43)	12.50 (6.88-18.07)
Max temperature of warmest month (Bio5; °C)	28.71 (22.58-35.76)	30.88 (22.58-37.34)
Mean temperature of wettest quarter (Bio8; °C)	18.15 (-1.52-26.70)	18.68 (-1.52-27.83)
Precipitation of driest quarter (Bio17; mm)	177.38 (35.40-344.38)	180.85 (19.21-424.48)
Precipitation of warmest quarter (Bio18; mm)	298.44 (188.31-482.10)	298.91 (104.20-628.75)
Mean elevation (m)	292.04 (0.45-1097.26)	294.43 (0.00-1529.13)
Percent of county forest land cover	36.01 (0.01-92.39)	30.56 (0.00-92.81)
Percent of county open space and low-intensity urban land cover	9.32 (0.84-49.51)	7.76 (0.25-64.17)

[†] States in CONUS that are split by or completely east of the 100th meridian (Fig. 1).

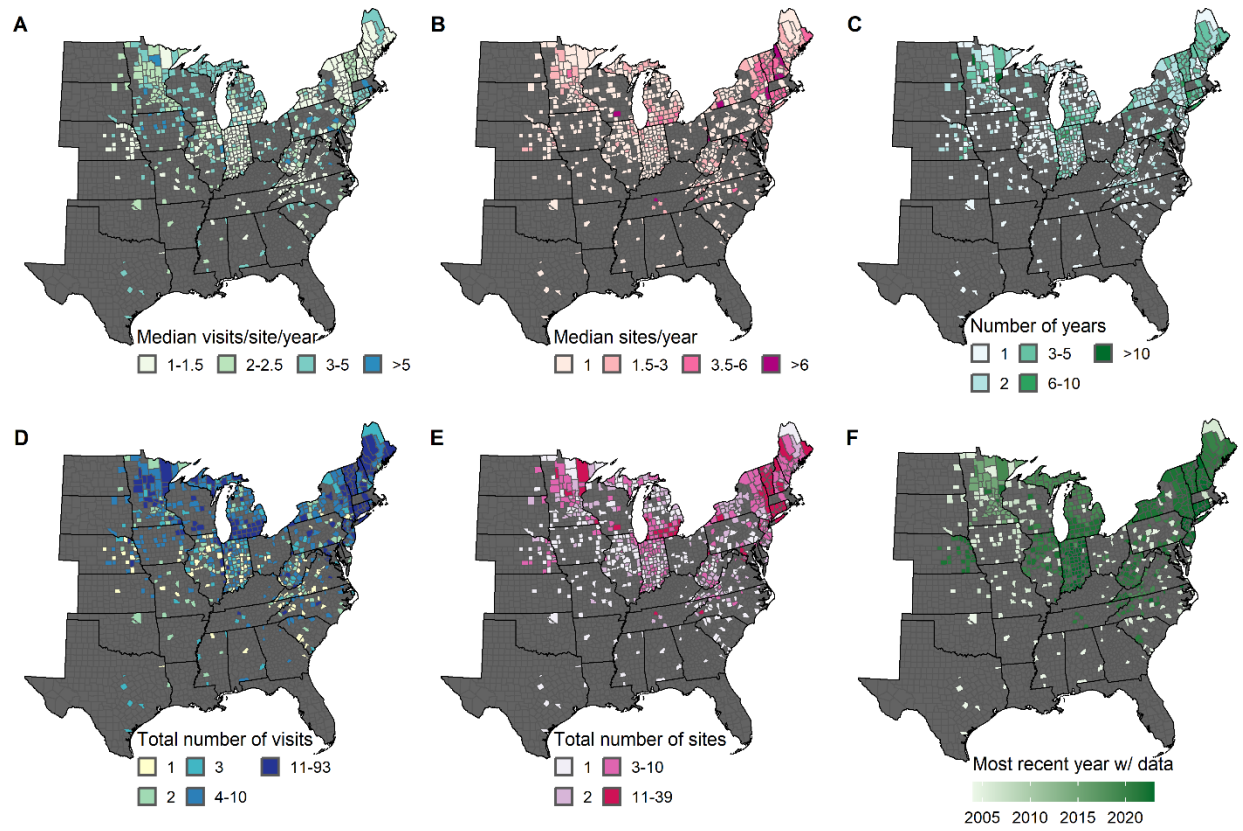


Figure A.1. Metrics of surveillance effort for data meeting our inclusion criteria per county. Sampling effort metrics include A) median collection visits per site per year, B) median unique sites per year, C) number of years with surveillance per county, D) total collection visits (2004-2023), E) total sites (2004-2023), and F) most recent year with included surveillance data. In all maps, dark grey counties either did not have tick surveillance data reported to the ArboNET Tick Module or data that did not meet our inclusion criteria.

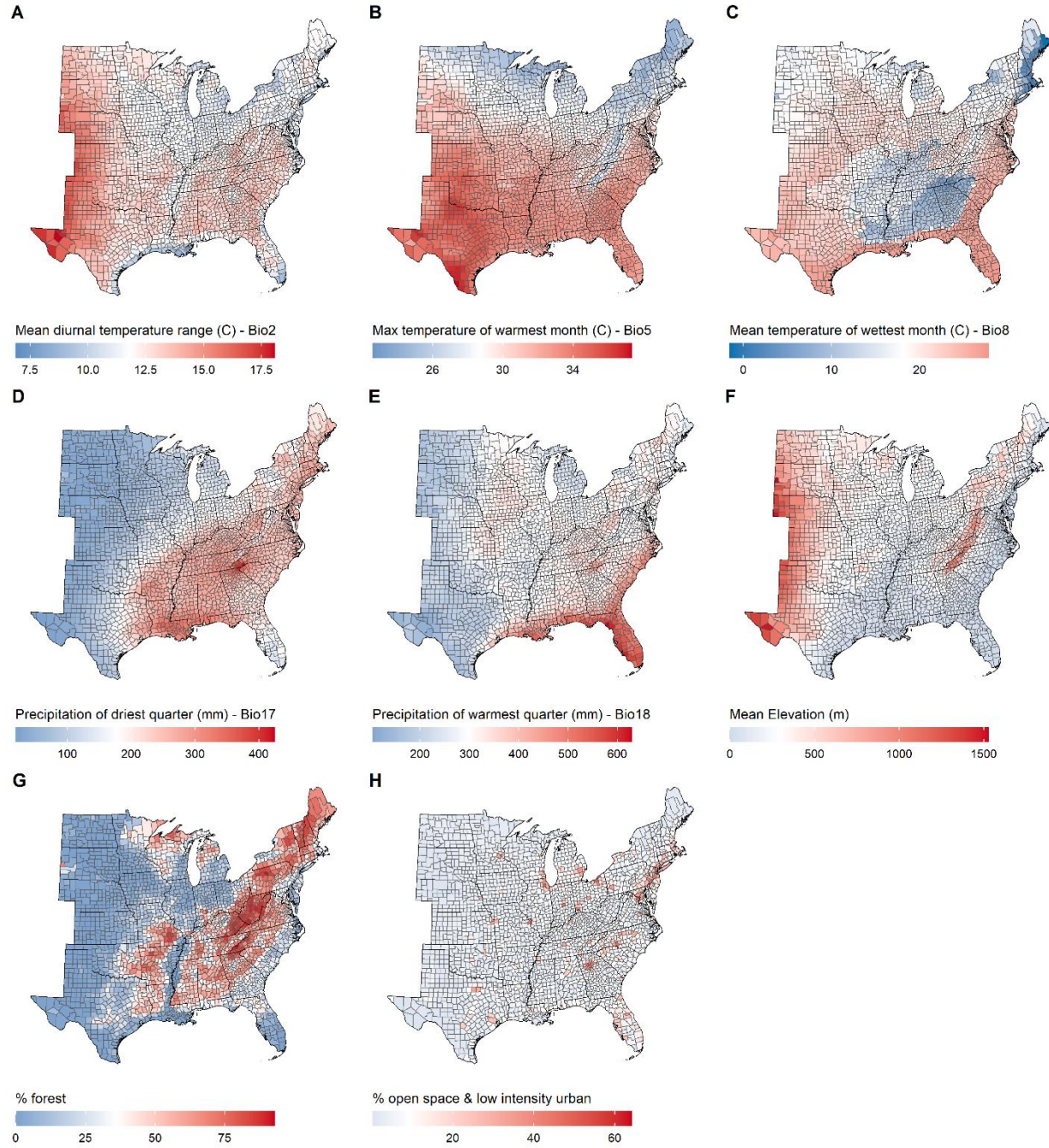


Figure A.2. Distribution of covariates considered for zero-inflated negative binomial (ZINB) model. The color ramps match the sign of the normalized covariate values used in model fitting. That is, white represents the mean value while blue and red represent values below and above the mean value, respectively. Normalization was based only on covariate values in counties used to fit the ZINB model. See Table A.3 for numerical details.

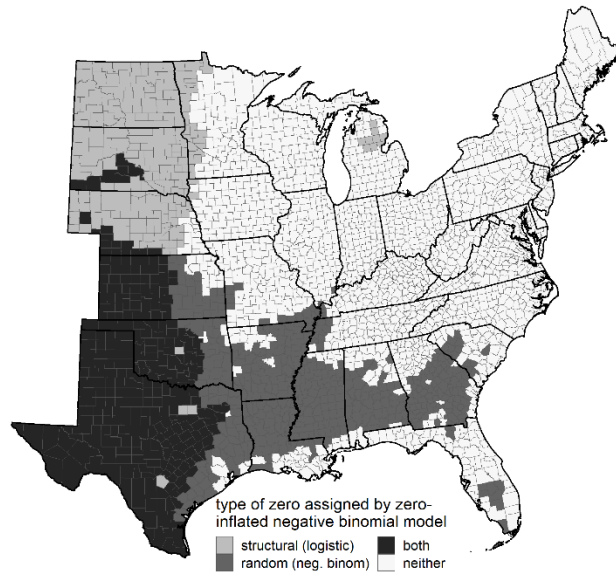


Figure A.3. Assignment of fitted zeros in zero-inflated negative binomial (ZINB) model. ZINB models account for observed zeros as either structural (zero-inflated piece; logistic model) or random (negative binomial piece). Assignment estimated for all counties in the eastern US using fitted coefficients for the logistic and negative binomial pieces separately. Counties for which both pieces assigned a zero are indicated as “both”. See Section 2.3 Model Fitting and Text A.1 for more details on ZINB models and Table 2 for fitted coefficients for each piece of the model.

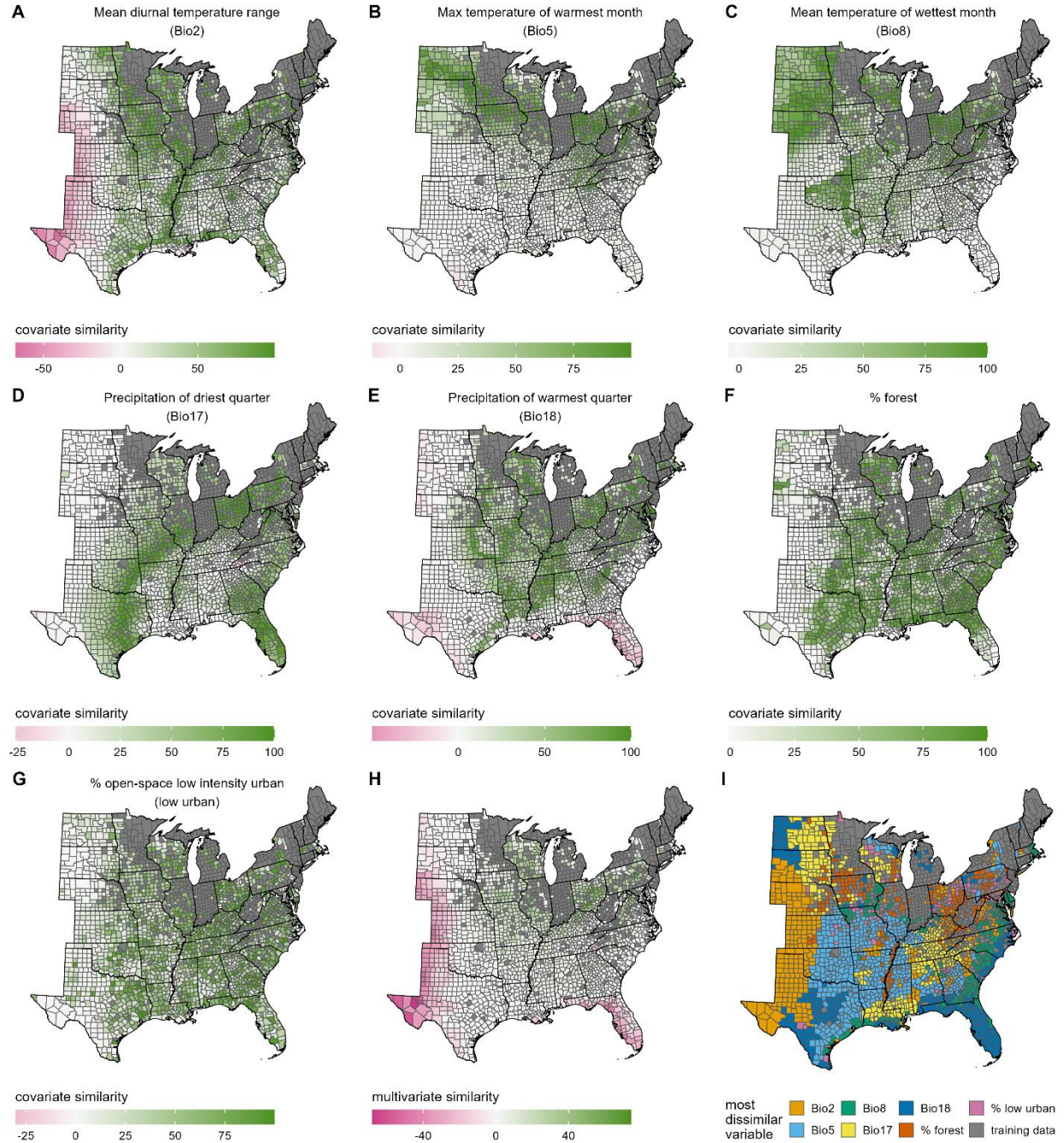


Figure A.4. Multivariate environmental similarity surface (MESS) for counties in the eastern US (projected set) relative to the counties with included tick surveillance data (grey counties; training data). A-G) Univariate similarity for each covariate included in the final ZINB model as measured by Euclidean distance from the center of the covariate distribution in the training data. H) Multivariate similarity as the minimum similarity across all variables. I) Covariate with the minimum similarity represented in H). In A-H), green shading (positive similarity) indicates the covariate falls within the distribution in the training data with darker shade representing closer to the center of the training data distribution. Pink shading (negative similarity) represents values outside the training distribution with darker shading representing larger distances and more dissimilar conditions. White (similarity = 0) indicates values at the min or max of the training distribution. See Fig. A.2. and Table A.3 for covariate values.

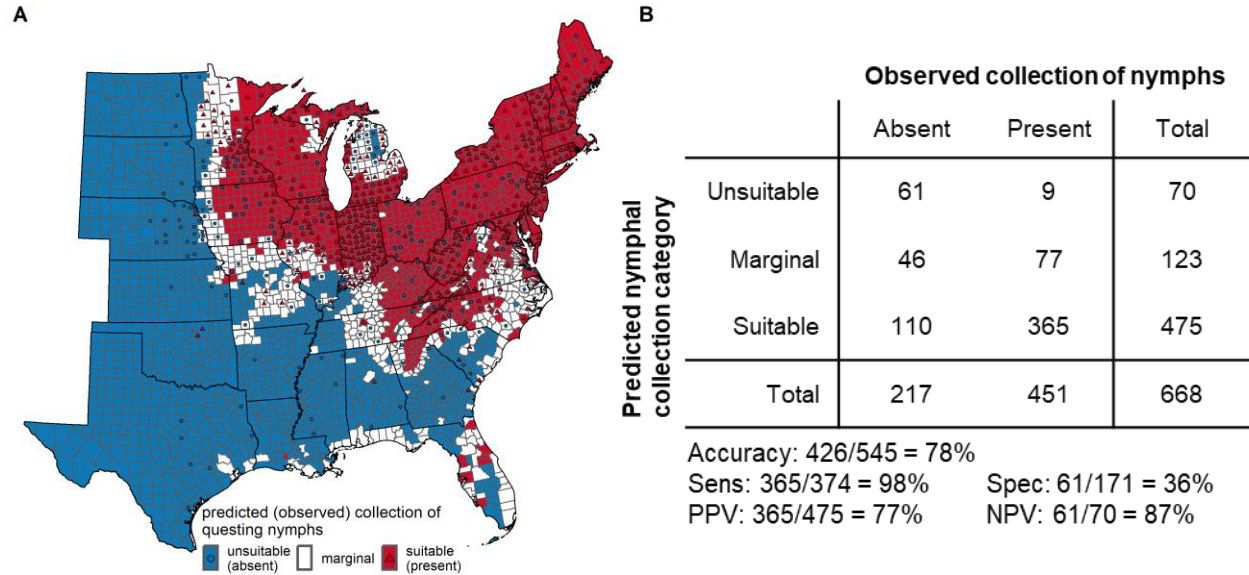


Figure A.5. Observed and predicted category of collection probability using optimal threshold identified with ROC curve analysis. A) Categorization of collection probability, accounting for the 95% CI, as unsuitable (<48.9% probability), marginal (95% CI contains 48.9% probability), or suitable (>48.9% probability), compared to observed collections (present or absence of collected nymphs). B) Contingency table summarizing observed and predicted collection probabilities presented in A). Diagnostic metrics below table based on counties with predictions of suitable or unsuitable. Sens: sensitivity; spec: specificity; PPV: positive predictive value; NPV: negative predictive value.

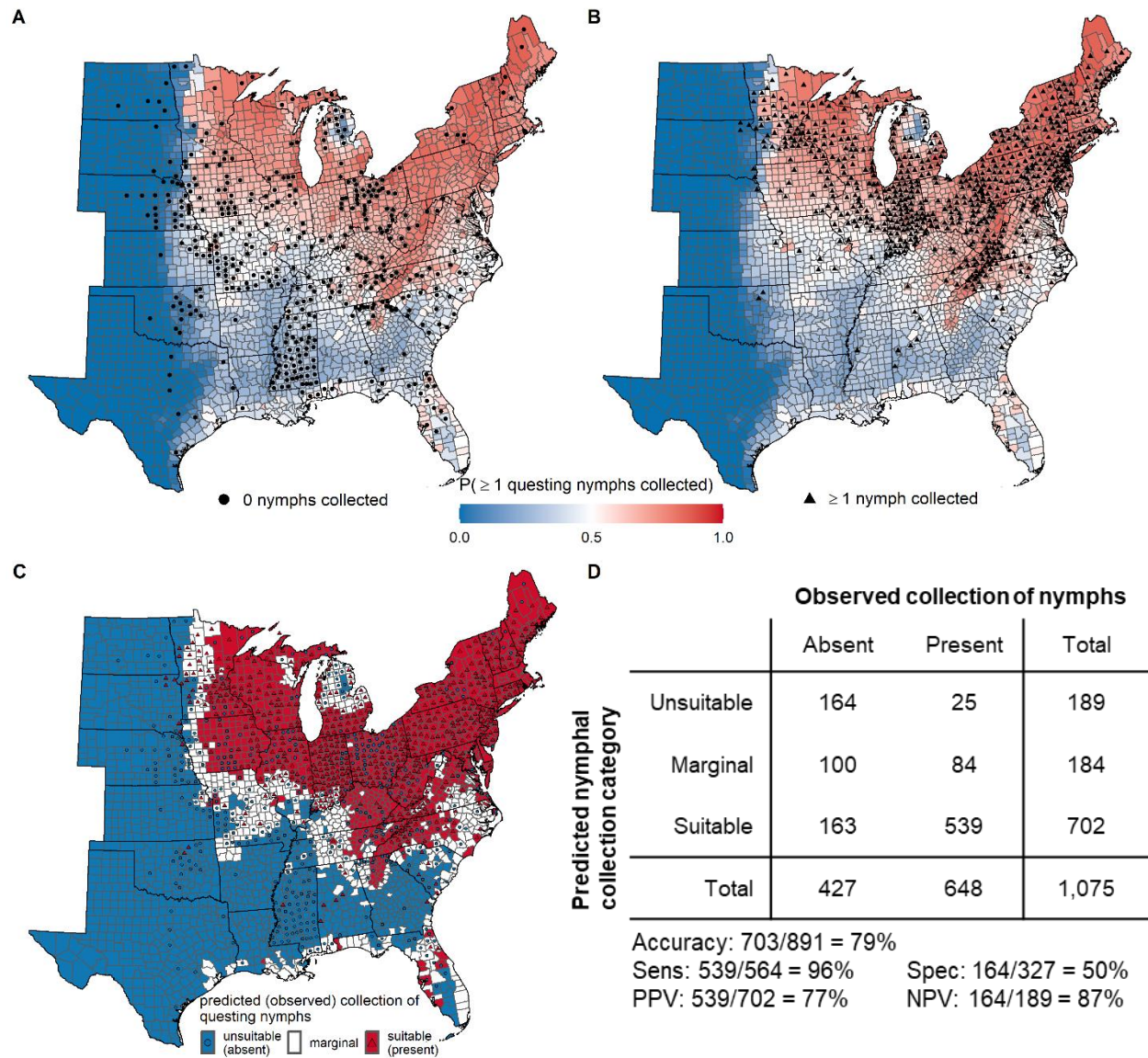


Figure A.6. Observed and predicted probability of collecting ≥ 1 questing nymph/1,000 m². Observed collections include full dataset of reported tick surveillance in the ArboNET Tick Module without exclusion criteria. County-level predicted collection probability across the eastern US with observed collections of A) zero (dots) or B) at least one (triangles) indicated. C) Categorization of collection probability, accounting for the 95% CI, as unsuitable ($<50\%$ probability), marginal (95% CI contains 50% probability), or suitable ($>50\%$ probability), compared to observed collections (present or absence of collected nymphs). D) Contingency table summarizing observed and predicted collection probabilities presented in C). Diagnostic metrics below table based on counties with predictions of suitable or unsuitable. Sens: sensitivity; spec: specificity; PPV: positive predictive value; NPV: negative predictive value.

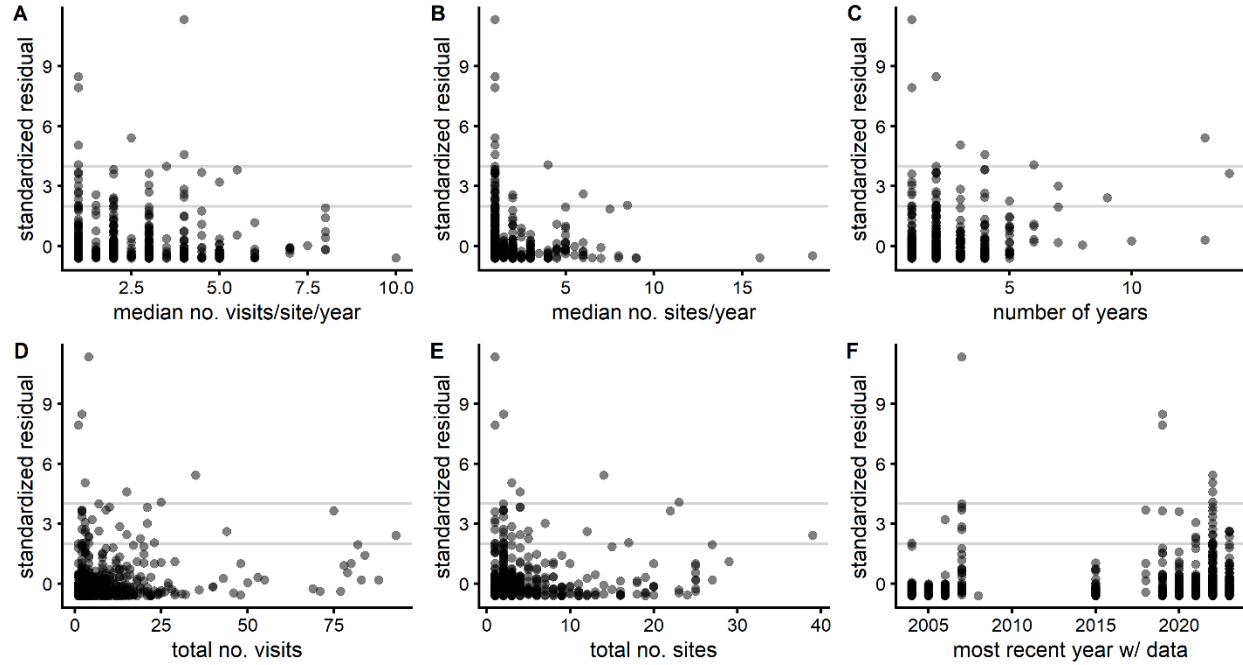


Figure A.7. Pearson's χ^2 standardized residuals for predicted density of *I. scapularis* nymphs per 1,000 m² per county relative to metrics of sampling effort: A) median collection visits per site per year, B) median unique sites per year, C) number of years with surveillance per county, D) total collection visits, E) total sites, and F) most recent year with included tick surveillance data. Horizontal lines delineate residuals with magnitude 2 ("high") and 4 ("extreme").

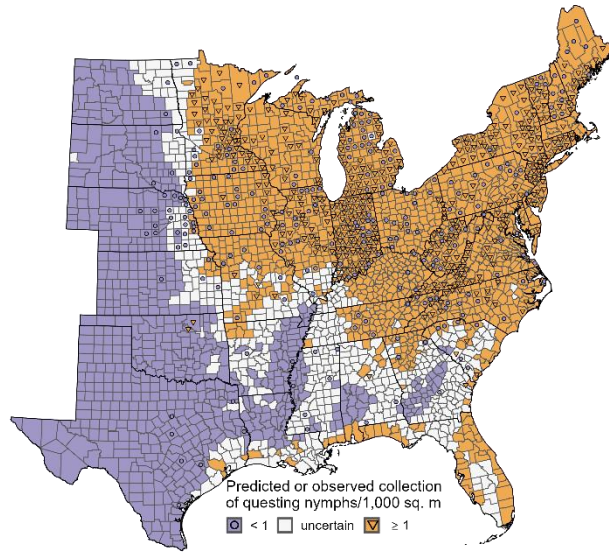


Figure A.8. Categorization of observed and predicted risk based on collections of ≥ 1 questing nymph per 1,000 m². Predictions from zero-inflated negative binomial model and account for 95% CI. Uncertain predictions (white) when 95% CI contains one while the 95% CI wholly below (purple) or above (orange) one, respectively. Observations from county peak DON derived from tick surveillance.

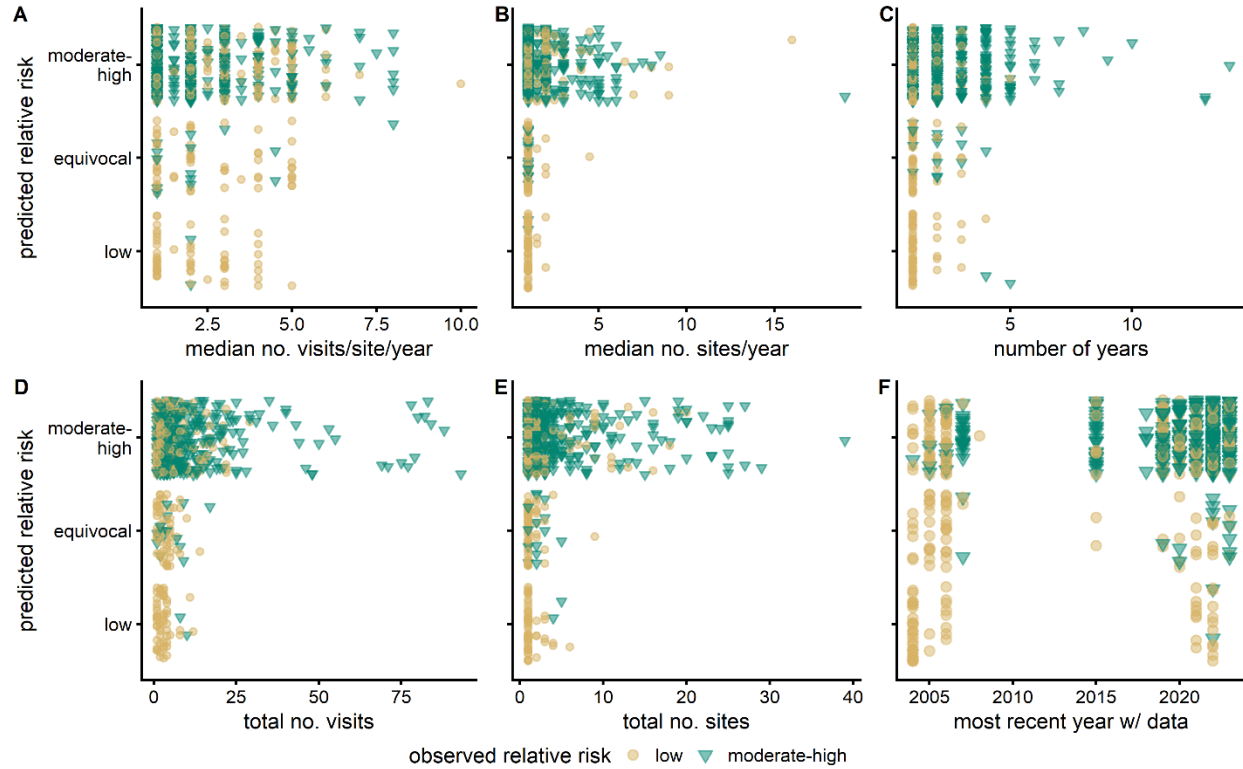


Figure A.9. Predicted relative risk category versus to metrics of sampling effort: A) median collection visits per site per year, B) median unique sites per year, C) number of years with surveillance per county, D) total collection visits, E) total sites, and F) most recent year with included tick surveillance data. Low risk category encompassed predictions for which the upper 95% CI bound < 2.2 nymphs/1,000 m². The moderate-high risk category encompassed predictions for which the lower 95% CI bound ≥ 2.2 nymphs/1,000 m². The equivocal risk category contained predictions for which the 95% CI contained values in both the low and moderate-high categories. Observed relative risk based on included tick surveillance data indicated by the shape and color of points. Points jittered y for visualization.

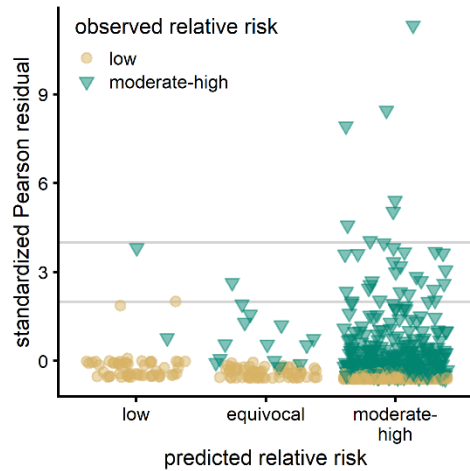


Figure A.10. Predicted relative risk versus Pearson’s χ^2 standardized residuals. Low risk category encompassed predictions for which the upper 95% CI bound < 2.2 nymphs/1,000 m². The moderate-high risk category encompassed predictions for which the lower 95% CI bound ≥ 2.2 nymphs/1,000 m². The equivocal risk category contained predictions for which the 95% CI contained values in both the low and moderate-high categories. Observed relative risk based on included tick surveillance data indicated by the shape and color of points. Points jittered slightly in x for visualization. Horizontal lines delineate residuals with magnitude 2 (“high”) and 4 (“extreme”).