HHS Public Access

Author manuscript

Ticks Tick Borne Dis. Author manuscript; available in PMC 2023 October 23.

Published in final edited form as:

Ticks Tick Borne Dis. 2022 September ; 13(5): 102000. doi:10.1016/j.ttbdis.2022.102000.

Predicting distributions of blacklegged ticks (*Ixodes scapularis*), Lyme disease spirochetes (*Borrelia burgdorferi* sensu stricto) and human Lyme disease cases in the eastern United States

James C. Burtis[*],
Erik Foster,
Amy M. Schwartz,
Kiersten J. Kugeler,
Sarah E. Maes,
Amy C. Fleshman,
Rebecca J. Eisen

Division of Vector-Borne Diseases, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Fort Collins, CO 80521, United States

Abstract

Lyme disease is the most commonly reported vector-borne disease in the United States (US), with approximately 300,000 -to- 40,000 cases reported annually. The blacklegged tick, *Ixodes scapularis*, is the primary vector of the Lyme disease-causing spirochete, *Borrelia burgdorferi* sensu stricto, in high incidence regions in the upper midwestern and northeastern US. Using county-level records of the presence of *I. scapularis* or presence of *B. burgdorferi* s.s. infected host-seeking *I. scapularis*, we generated habitat suitability consensus maps based on an ensemble of statistical models for both acarological risk metrics. Overall accuracy of these suitability models was high (AUC = 0.76 for *I. scapularis* and 0.86 for *B. burgdorferi* s.s. infected-*I. scapularis*). We sought to compare which acarological risk metric best described the distribution of counties reporting high Lyme disease incidence (≥10 confirmed cases/100,000 population) by setting the models to a fixed omission rate (10%). We compared the percent of high incidence counties correctly classified by the two models. The *I. scapularis* consensus map correctly classified 53% of high and low incidence counties, while the *B. burgdorferi* s.s. infected-*I. scapularis* consensus map classified 83% correctly. Counties classified as suitable by the *B. burgdorferi* s.s. map showed a 91% overlap with high Lyme disease incidence counties with over a 38-fold difference in Lyme disease incidence between high- and low-suitability counties. A total of 288 counties were classified as highly suitable for *B. burgdorferi* s.s., but lacked records of infected-*I. scapularis* and

[*]Corresponding author. ptd6@cdc.gov (J.C. Burtis).

CRediT authorship contribution statement

James C. Burtis: Conceptualization, Methodology, Formal analysis, Visualization, Writing – original draft. **Erik Foster:** Data curation, Validation. **Amy M. Schwartz:** Data curation, Validation. **Kiersten J. Kugeler:** Data curation, Validation, Writing – review & editing. **Sarah E. Maes:** Data curation. **Amy C. Fleshman:** Data curation. **Rebecca J. Eisen:** Conceptualization, Methodology, Writing – review & editing.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ttbdis.2022.102000.

were not classified as high incidence. These counties were considered to represent a leading edge for *B. burgdorferi* s.s. infection in ticks and humans. They clustered in Illinois, Indiana, Michigan, and Ohio. This information can aid in targeting tick surveillance and prevention education efforts in counties where Lyme disease risk may increase in the future.

1. Introduction

Lyme disease is the most commonly reported vector-borne disease in the United States (US), with approximately 30,000 -to- 40,000 estimated cases occurring annually (Schwartz et al., 2017). The number of counties considered high incidence for Lyme disease has expanded over time (Kugeler et al., 2015), and case numbers have increased most rapidly in counties where Lyme disease has recently been established (Burtis et al., 2016). The increase in cases and geographic expansion of Lyme disease in recent decades is driven in part by the range expansion of *Ixodes scapularis*, the primary vector of *Borrelia burgdorferi* sensu stricto in the eastern US (the primary causative agent of Lyme disease, referred to as *B. burgdorferi* hereafter) (Eisen et al., 2016).

The majority of high-incidence counties are in the upper midwestern and northeastern states, but suitable habitat for *I. scapularis* has been predicted to cover much of the eastern US (Hahn et al., 2016, 2017; Peterson and Raghavan, 2017), including counties in the southeast where Lyme disease incidence is low (Schwartz et al., 2017). Previous studies suggest that the density of host-seeking *B. burgdorferi*-infected nymphs (DIN) is a better predictor of Lyme disease case occurrence than vector presence or abundance data alone (Eisen and Eisen, 2016). Estimation of DIN requires that both tick density and pathogen infection prevalence are collected. The Centers for Disease Control and Prevention (CDC) guidance suggests a minimum of 750 m of dragging distance to estimate density, and at least 25 ticks be tested for infection prevalence estimates (CDC, 2018). As a result, such estimates are costly. Due to a lack of systematic tick-based surveillance efforts and the cost of generating such data, DIN estimates are lacking for most counties in the US.

In 2018 CDC initiated a national tick-based surveillance program (CDC, 2018; Eisen and Paddock, 2021) to provide current and accurate information to the public, clinicians, and policy makers regarding the distribution and abundance of medically important ticks and the distribution and prevalence of their associated human pathogens. Data generated by this program, coupled with a review of historical data in the peer-reviewed literature, yielded recent county-scale updates for the geographic distribution of *I. scapularis* and *B. burgdorferi* in the US (Eisen et al., 2016; Fleshman et al., 2021, 2022). The reported county-level distribution of host-seeking *I. scapularis* infected with *B. burgdorferi* is more constrained than the distribution of *I. scapularis* (Fleshman et al., 2021, 2022). However, the lack of systematic sampling and tick testing efforts across the US results in an under-representation of the actual distribution of infected ticks. Habitat suitability models can aid in efforts to identify the range of vectors and their associated pathogens, which may be under-represented by current tick-based surveillance practices. While predictions based upon habitat suitability models are reliant upon the extent of the underlying dataset, these models may be helpful in identifying locations with suitable habitat that lack surveillance

data. Habitat suitability models have been used in other biological systems to focus sampling efforts for both rare and invasive species (Crall et al., 2013; Aizpurua et al., 2015) and may help to focus tick collection and testing efforts in the eastern US.

In this study, we developed separate habitat suitability maps to estimate the predicted suitable range of counties where *I. scapularis* is likely to be established and where *B. burgdorferi*-infected host-seeking *I. scapularis* are likely to occur. We then used the resulting models to assess which acarological risk metric best predicts the distribution of counties considered to be high incidence for Lyme disease. Next, we projected the acarological risk models to identify potential 'leading edge' counties where infected *I. scapularis* populations may be expanding and Lyme disease cases may be increasing. This information will be useful in determining the geographic extent of suitable habitat for *B. burgdorferi*-infected *I. scapularis*. It will also aid in identifying counties in the eastern US where tick surveillance efforts should be focused and where acarological risk is likely to increase in the future.

2. Methods

2.1. Field data for acarological risk metrics

County level datasets were generated for two different acarological risk metrics. The first included observed presence of *I. scapularis* of any life stage and the second included observations of *B. burgdorferi*-infected host-seeking *I. scapularis*. These will be referred to as "*I. scapularis*" or "*B. burgdorferi*" models hereafter, respectively.

For the *I. scapularis* dataset, observations were derived from a combination of published data for the distribution of *I. scapularis* reported by Eisen et al. (2016) and data submitted to the ArboNET Tick Module through 2020. The ArboNET Tick Module is a data portal maintained by CDC that allows public health agencies to report surveillance data on the presence and abundance of medically important ticks and the presence and prevalence of their associated human pathogens in ticks. In previous studies (Dennis et al., 1998; Eisen et al., 2016) counties were classified as 'established' if six or more ticks, or more than one life stage were collected in a single year; counties were classified as 'reported' if these thresholds were not reached, but at least one tick of any life stage was detected. The 'reported' counties were not included to reduce the potential for false positives in counties lacking an established population, potentially representing counties where individual *I. scapularis* were transported long distances by hosts (Schneider et al., 2015; Scott, 2016). This also allowed for direct comparisons against previous *I. scapularis* modeling efforts by Hahn et al. (2016, 2017). Counties where no *I. scapularis* were collected were classified as 'no records.' In our analysis, 'established' counties were coded as "established' (1), while 'reported' and 'no records' counties were coded as 'not established' (0). A total of 1001 counties were coded as 'established', predominantly in the Northeast, Upper Midwest and along the Atlantic coast.

Observations of *B. burgdorferi*-infected *I. scapularis* were derived from a combination of published historical data from the literature and ArboNET Tick Module data through the end of 2020. Testing methods conformed to CDC tick surveillance guidance (CDC, 2018) and the majority of ticks submitted to CDC were tested using a series of real-time polymerase

chain reaction assays described in Graham et al. (2018). The geographic distribution of these data were published previously (Fleshman et al., 2021). Counties where one or more host-seeking *B. burgdorferi*-infected *I. scapularis* was collected were coded as 'present' (1), while counties without record of infection were coded as 'no records' (0). For *B. burgdorferi* presence, we required only a single infected tick to be detected, rather than the six required for *I. scapularis* establishment, due to the additional sampling effort required to collect infection data; increasing the number of ticks required to be infected would penalize counties with low prevalence or low tick abundance. A total of 402 counties were coded as 'present', predominantly in the Northeast and Upper Midwest.

2.2. Human Lyme disease incidence data

Confirmed cases of Lyme disease voluntarily reported to CDC from state and local health departments via the National Notifiable Diseases Surveillance System (NNDSS) between 2000 and 2019 were collated by county of residence across the eastern US (CDC, 2021a). The surveillance case definition for confirmed Lyme disease was modified during this 20-year period. Between 2000 and 2007 cases were considered confirmed if either erythema migrans (EM) rash was present, or specific late manifestations affecting the musculoskeletal, nervous, or cardiovascular systems were present and associated with positive serological laboratory evidence. Beginning in 2008, laboratory criteria were modified to increase specificity (CSTE, 2007). Acceptable laboratory evidence included: (1) a positive culture for *B. burgdorferi*, (2) two-tier testing interpreted using established criteria, or (3) single-tier IgG immunoblot seropositivity interpreted using established criteria. Beginning in 2017, laboratory evidence of infection was required for all cases from low-incidence states, including for those associated with EM rash (CDC, 2021b). The annual incidence per 100,000 persons between 2000 and 2019 was calculated using the 2010 census county population estimates (US Census, 2021). The national Lyme disease case data we used are reliant upon the ability of each state to report data to CDC, which can vary between years. Overall, under-reporting is more likely in high-incidence areas, while over-reporting is more likely in areas where incidence is low (CDC, 2021a). To account for this, counties were considered 'high-incidence' if there were ≥10 confirmed cases per 100, 000 persons for three or more consecutive years between 2000 and 2019. Our definition focuses on counties consistently reporting high incidence of Lyme disease, rather than single year reports of pure incidence values or cumulative incidence over the entire observation period. This was done to ensure capture of counties where high-risk Lyme disease emerged more recently (Kugeler et al., 2015), and to minimize potential of travel-associated cases to lead to misclassification of county status (Forrester et al., 2015).

2.3. Climate and landscape predictors

We selected a variety of climate and landscape predictors to construct our habitat suitability models based upon the ecology of *I. scapularis* and potential to affect *B. burgdorferi* transmission dynamics. We included 19 climate predictors from WorldClim, which are commonly used for bioclimatic modeling (Hijmans et al., 2005). The original dataset uses weather observations between 1950 and 2000 to generate climate layers at a 1 km × 1 km spatial resolution. We updated these climate layers using Daymet data (Thornton et al., 1997, 2016) with records collected between 1980–2015 as previously described in

Johnson et al. (2017). Three additional climate variables were also evaluated. Average monthly growing degree days above 10 °C (GDD10) between December and February was included as a measure of sustained cold weather that might affect development of overwintering *I. scapularis* (Ogden et al., 2004). Vapor pressure between March and June was included as a measure of water vapor in the air during a period when *I. scapularis* nymphs likely experience high mortality rates (Burtis et al., 2019). Finally, snow water equivalent between November and April was included as a measure of snowfall over the winter which may insulate overwintering ticks (Linske et al., 2019). We also evaluated two landscape predictors, percent forest cover based upon the 2006 national land cover database (Fry et al., 2011) and elevation based upon global multi-resolution terrain elevation data from 2010 (Danielson and Gesch, 2011), both of which are known to be significant predictors of *I. scapularis* presence in regional models (Estrada-Peña, 2002; Diuk--Wasser et al., 2010; Hahn et al., 2016). The Zonal Statistics tool was used in QGIS (v. 3.14.1) to generate county-level estimates for the eastern US. Minimum, maximum, or mean values were used to generate layers for different predictors. This was done to explore the effect of temperature and precipitation extremes. Additional details regarding the predictors are described in Table 1.

To select predictors, we first sorted them by percent deviance explained, and eliminated those that explained <5% of the deviance. Percent deviance is a goodness of fit statistic, similar to an r-squared value (Talbert and Talbert, 2001; Guisan et al., 2017). Many climate predictors were collinear, so a correlation matrix of all predictors was created prior to their use in the models described below. Spearman Rho correlations against both the county-level acarological data (i.e. observed presence of *I. scapularis* or *B. burgdorferi*) and each other were generated. Predictors were further sorted in descending order according to their spearman Rho correlation scores against the county-level acarological data. If two predictors were correlated with one another (> 0.80) the predictor that was most strongly correlated with the county-level acarological data was retained while the other was eliminated. This same process was repeated for both the *I. scapularis* and *B. burgdorferi* datasets.

2.4. Acarological risk modeling

We generated separate models based upon the two datasets (*I. scapularis* and *B. burgdorferi*), but the same process was followed and presence at the county level was coded as described above. The extent of both datasets was limited to states in the eastern US, representing the expected range of *I. scapularis* (Diuk-Wasser et al., 2010; Hahn et al., 2016) (Figs. 1 and 2). Five modeling algorithms were used for each of the two acarological outcomes: (1) boosted regression tree (BRT), (2) generalized linear model (GLM), (3) maximum entropy (Maxent), (4) multivariate adaptive regression splines (MARS), and (5) random forest (RF) (Talbert and Talbert, 2001).

BRT is a boosting approach that iteratively fits trees, first generating a sequence of simple trees and then creating new trees based upon the residuals of those already created. It is a flexible approach that can account for complex relationships between predictor and response variables (Elith et al., 2008; Merow et al., 2014). GLM is a generalized ordinary least squares regression which links to a binomial function. This approach is suitable when

relationships between predictor and response variables are not overly complex (Guisan et al., 2017). Maxent is a maximum entropy approach which generates probability distributions to match the dataset, with the distribution which maximizes uncertainty (entropy) being considered the most suitable. It is highly flexible and particularly well-suited for presence-only datasets (Elith et al., 2011). MARS represents a more flexible regression which allows relationships between predictor and response variables not to conform to a predefined shape. This allows for the analysis of complex relationships between predictors and responses (Guisan et al., 2017). RF is a bagging or bootstrapping aggregation approach wherein trees are fit to different bootstrapped subsamples of the data and the averaged probability across all runs is extracted (Breiman, 2001). Each modeling approach has different underlying assumptions, which can affect their outcome. Using an ensemble approach can help to account for differences in these assumptions and their effect on the model output. The models produced probability scores, referred to as suitability scores hereafter, for each county. Suitability scores were converted to binary data (high or low suitability) by using different probability thresholds, as described in the next section.

We used a 10-fold cross-validation method to generate performance statistics for each model. The training data were divided into 10 equal subsets and the models were run 10 separate times, excluding one subset each time. These are referred to as the ‘testing’ runs hereafter, while the model ‘training’ runs use the entire dataset without excluding data points. The Receiver Operating Characteristic (ROC) curve and the resulting area under the curve (AUC) were assessed for the training and testing runs. The ROC curve is a plot of the true positive rate against the false positive rate at different suitability thresholds. The AUC, derived from the ROC, is a threshold unbiased measure of model accuracy. A value of 1 indicates a ‘perfect’ model and values ≤ 0.5 indicate a poor distinction between counties classified as high- or low-suitability (Fielding and Bell, 1997). Overfitting was assessed by comparing the AUC values for the training and test runs. If AUC values differed by >0.05 the model was considered overfit (Springer et al., 2015; Hahn et al., 2016). We also checked for large differences in specificity, sensitivity, positive predictive value, negative predictive value, percent deviance explained, percent correctly classified, and the correlation coefficient between the training and testing runs (Elith et al., 2008). The percent correctly classified (PCC) is a measure of overall model accuracy and is calculated:

$$PCC = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Nagative}$$

The correlation coefficient represents the overall linear relationship between the field data and the model output. The RF and BRT models were overfit and therefore not used in the consensus maps described below. All models were generated using the VisTrails Software for Assisted Habitat Modeling (SAHM v. 2.2.3) using the default model settings.

2.5. Model thresholding and visualization of consensus maps

Using the output for the modeling algorithms (GLM, MARS, Maxent), we generated continuous suitability scores and converted scores to binary outcomes (high or low suitability) based on the following thresholding criteria. We dichotomized the model using

a threshold suitability score that simultaneously maximized the sum of sensitivity and specificity as ascertained using the ROC plots. This is expected to yield the most constrained geographic distribution of highly suitable counties when using a presence-only dataset (Liu et al., 2013; Hahn et al., 2017). Recognizing we have greater confidence in presence than pseudo-absence data, we also dichotomized the models based on fixed omission error (false negative) rates (10% and 5%) showing sensitivities of 90 or 95%. Since we have confidence in the presence data, we used the false negative rate to standardize error rate against those observations rather than the pseudo-absence data (Peterson, 2014).

We overlaid each thresholded binary statistical model to create two separate consensus maps of suitability, one for *I. scapularis* and another for *B. burgdorferi*. In consensus maps, counties that the majority (≥2) of models classified as highly suitable maintained their classification. All other counties that one or zero models predicted to be suitable were classified as low-suitability. Map performance was evaluated by comparing suitability with observed presence of *I. scapularis* and *B. burgdorferi* to derive the sensitivity, specificity, positive predictive value, negative predictive value, and percent correctly classified. The binary model output was combined to create consensus maps using QGIS v. 3.14.1 (QGIS Development Team, 2021).

2.6. Comparison of acarological risk models for classifying high incidence Lyme disease counties

We sought to determine which acarological risk metric (*I. scapularis* or *B. burgdorferi*) better described the reported distribution of counties classified as high incidence for Lyme disease. We normalized outcomes of individual *I. scapularis* and *B. burgdorferi* statistical models (GLM, MARS, and Maxent) by thresholding the models at 90% sensitivity. Recognizing that tick-based surveillance is not conducted evenly or randomly across the eastern US, we fixed the omission rate evenly at 10% for both the *I. scapularis* and *B. burgdorferi* consensus maps. The two consensus maps were evaluated using three criteria: (1) overall accuracy and fit of the models to predict counties classified as high incidence for Lyme disease based upon the AUC and Akaike information criterion (AIC), respectively, derived from logistic regressions, (2) discrimination in Lyme disease incidence between high- and low-suitability counties, and (3) percentage of high-incidence counties correctly classified.

For the logistic regressions, we used the binary output from each of the two consensus maps (*I. scapularis* or *B. burgdorferi*) as predictors of high-incidence counties, also coded as binary (high incidence or not). We conducted two logistic regressions to generate AIC values. A difference of ≥2 in the AIC score indicate significant model improvement, with the lower value indicating a superior model (Burnham and Anderson, 2004). The AUC values were generated using the pROC package in the R statistical environment (v. 4.0.3). To determine discrimination between high- and low-incidence counties, we subtracted the median Lyme disease incidence in counties classified as low-suitability from the median incidence in high-suitability counties. The median Lyme disease incidence was also compared between counties classified as high- and low-suitability by both consensus maps. The distribution of Lyme disease incidence data was not normal, so Wilcoxon rank sum tests were used for these comparisons. The alignment between counties classified as

high-suitability by the two consensus maps and high-incidence counties was determined by calculating the percentage of counties correctly classified as high- or low-suitability. The consensus map based upon the 90% sensitivity models that most accurately classified high Lyme disease incidence counties according to the criteria above was expanded to 95% sensitivity to identify additional potential 'leading edge' counties where acarological risk and Lyme disease incidence are predicted to increase. All statistical analyses were conducted in the R statistical environment v. 4.0.3 (R Core Team, 2021).

3. Results

3.1. Variable selection and model performance

Using the variable selection method described, we selected six predictors to model suitability for *I. scapularis* and seven to model suitability for *B. burgdorferi* (Table 3). Both the RF and BRT modeling algorithms exhibited overfitting. The BRT models for both datasets had differences in AUC > 0.05 between the testing and training models. The RF models exhibited low sensitivities (<70%) and differences in sensitivities and specificities were >10% between testing and training runs for both datasets. Furthermore, upon visual inspection of both the RF and BRT output these models showed strong overfitting tendencies, almost entirely predicting presence only in counties where *I. scapularis* or *B. burgdorferi* had been observed. Adjusting model parameters did not improve the results over default settings. Hence, only the GLM, MARS, and Maxent modeling algorithms were used to generate the two consensus maps. The full output for the testing and training runs of these three models is shown in Table 2. The Maxent modeling algorithm does not have an internal variable selection process, so retained all predictors. Predictor variables that do not explain a significant amount of variation have low normalized contribution values in the Maxent models.

4. Predictor response curves

4.1. I. scapularis model response curves

For the *I. scapularis* dataset, all six predictors, BIO2, BIO5, BIO18, BIO19, percent forest cover, and elevation, were retained by each of the three modeling algorithms (see maps in Supplemental Fig. S1). Each predictor exhibited a different relationship with the habitat suitability score. The mean diurnal temperature range (BIO2), showed a negative relationship with the suitability score for *I. scapularis*. Lower diurnal ranges had higher suitability scores. The maximum temperature during the warmest month (BIO5) showed an s-shaped relationship with the suitability score, with the highest suitability scores at low temperatures. Amount of precipitation in the warmest quarter (BIO18) showed an approximately linear positive relationship. Counties with more summer precipitation had higher suitability scores. Precipitation of the coldest quarter (BIO19) showed a negative relationship with the lowest values having the highest suitability scores. Percent forest cover showed a positive relationship with the suitability score up to approximately 75% forest cover above which point the relationship became negative. Elevation showed a negative relationship with high elevations being the least suitable (Fig. 3).

4.2. B. burgdorferi model response curves

For the *B. burgdorferi* dataset, all predictors were used in each modeling algorithm, with the exception of BIO18 which was dropped from the MARS model (see maps in Supplemental Fig. S2). Four predictors, BIO5, BIO18, BIO19, and percent forest cover, showed unimodal relationships with the suitability score being highest at intermediate values. The mean diurnal temperature (BIO2) showed a similar pattern to that of the *I. scapularis* dataset, with the highest suitability scores at temperatures between approximately 14 and 18 °C. The effect of isothermality (BIO3) was slightly inconsistent between modeling algorithms. The GLM showed a curved increasing relationship with the highest suitability score at high isothermality, the MARS model showed a relatively flat relationship, with a slight increase in the suitability score as isothermality increases. The Maxent model showed a relatively flat relationship with a slight peak in suitability at intermediate isothermality. The effect of the mean temperature of the warmest quarter (BIO8) also showed some variation between modeling algorithms. The GLM model showed a slight increase in the suitability score at intermediate temperatures after which the relationship was flat, while the Maxent and MARS models showed the same small increase in the suitability score at low temperatures followed by a steep increase at higher temperatures. There was a negative relationship between the mean temperature of the driest quarter and suitability with the warmest temperature being least suitable (Fig. 3). GLM models can struggle to account for complex relationships between response and predictor variable, but the majority of predictor relationships are linear, unimodal, or s-shaped and align well with the output from the MARS and Maxent models, indicating that the GLM can appropriately model these relationships (Guisan et al., 2017). This is the case for both the *I. scapularis* and *B. burgdorferi* GLM models.

4.3. Consensus map performance based on comparisons with I. scapularis or B. burgdorferi presence records

4.3.1. I. Scapularis model performance—There was strong agreement in counties predicted to be high-suitability between the three modeling algorithms using the *I. scapularis* dataset. With the sum of sensitivity and specificity maximized, the resulting consensus map predicts most counties in the Northeast, Mid-Atlantic, and Upper Midwest to be highly suitable. Much of the coastal Southeast is also predicted to be highly suitable, along with a portion of the northern Appalachian Mountain range in West Virginia. The predicted high-suitability area expands inland in the southeast as model sensitivity is increased and specificity is decreased (Fig. 1). The consensus map yielded an overall accuracy (AUC) of 0.76. The *I. scapularis* model based upon models with the sum of sensitivity and specificity maximized had the highest specificity and percent correctly classified values. The sensitivity of the *I. scapularis* consensus map with the sum of sensitivity and specificity maximized was 77%. That is, 77% of counties where *I. scapularis* was classified as established were considered highly suitable by the model. When model sensitivity was increased to 90%, specificity was reduced to 54%. In other words, 54% of counties lacking records of *I. scapularis* establishment were classified as low-suitability. Negative predictive values (i.e., the model predicted a county to be low-suitability and tick records were lacking for that county) were high (>80%) at all sensitivity settings. Positive predictive values (i.e., counties

classified as highly suitable where ticks were recorded as established) were overall low, with a maximum of 65% (Table 4).

4.4. B. burgdorferi model performance

There was strong agreement between the three modeling algorithms based upon the *B. burgdorferi* dataset, which yielded an overall accuracy (AUC) of 0.86. With the sum of sensitivity and specificity maximized, the high-suitability counties predicted by this consensus map are constrained primarily to the Northeast, Mid-Atlantic, and Upper Midwest. The number of high-suitability counties expanded with increasing model sensitivity. In contrast to the *I. scapularis* model, most of the Southeast was classified as low-suitability (Fig. 2). The consensus map based upon models with the sum of sensitivity and specificity maximized had the highest specificity and percent correctly classified values. The sensitivity of the *B. burgdorferi* consensus map with the sum of sensitivity and specificity maximized was 91%, so the effect of increasing model sensitivity to 90% was limited only increasing specificity by 0.3%. Negative predictive values were high (>97%) at all sensitivity settings. That is, more than 97% of counties classified as low-suitability lacked records of infected ticks. Positive predictive values were overall low, with a maximum of 45.7% of counties classified as highly suitable for *B. burgdorferi* also having records of *I. scapularis*-infected with *B. burgdorferi*. The low positive predictive values observed with both the *B. burgdorferi* and *I. scapularis* consensus models are consistent with under-reporting, which may have resulted from inconsistent geographic coverage of the surveillance program, range expansion, or a combination of both. Overall, the percentage of correctly classified counties was higher for the *B. burgdorferi* consensus map than that for *I. scapularis* (Table 4).

4.5. Relationship between consensus maps and Lyme disease incidence

4.5.1. I. Scapularis consensus map and Lyme disease incidence—

The AIC value of the logistic regression comparing the binary output from the *I. scapularis* consensus map with models set to 90% sensitivity against high-incidence counties was 2958 and the AUC value was 0.72. The median Lyme disease incidence in high-suitability counties by this consensus map [0.67 cases per 100,000 persons (IQR: 0.15–6.20)] was significantly higher than incidence in low-suitability counties [0.06 cases per 100,000 persons (IQR: 0.00–0.39) ($z = -21.47$, $p < 0.001$) (Table 5). The overall percentage of high- and low-incidence counties correctly classified by the *I. scapularis* map was 53.4%. The sensitivity was 99.3%, specificity was 44.5%, positive predictive value was 25.7%, and negative predictive value was 99.7% (Table 6). Many counties in the coastal Southeast that are not high-incidence were classified as highly suitable by the *I. scapularis* consensus map, yielding a relatively high false positive rate.

4.6. B. burgdorferi consensus map and Lyme disease incidence

The AIC value of the logistic regression comparing the binary output from the *B. burgdorferi* consensus map with models set to 90% sensitivity against high-incidence counties was 1504 and the AUC value was 0.85. The AIC reflects a better fit model compared with the *I. scapularis* model (ΔAIC = 1454). Comparison of AUC values indicate

that the *B. burgdorferi* consensus map better predicts high and low Lyme disease incidence counties than the *I. scapularis* consensus map. The median Lyme disease incidence of high-suitability counties according to the *B. burgdorferi* consensus map with models set to 90% sensitivity [5.71 cases per 100,000 persons (IQR: 0.87–30.85)] was significantly higher than incidence in counties classified as low-suitability [0.15 cases per 100,000 persons (IQR: 0.00–0.49)] ($z = -32.72$, $p < 0.001$) (Table 5). The *B. burgdorferi* consensus map shows better discrimination in Lyme disease incidence between high and low-suitability counties compared to the map based upon the *I. scapularis* dataset. The percent of high-incidence counties correctly classified as high- or low-incidence (high- or low-suitability, respectively) by the *B. burgdorferi* consensus map was 83.8%. The model classified 91.8% of high-incidence counties as high-suitability (sensitivity), and 82.3% of low-incidence counties as low-suitability (specificity). Overall, 50.2% of counties classified as suitable reported a high incidence of Lyme disease (PPV = 50.2). The model yielded a low false negative rate, with 1.34% of counties classified as low-suitability where high incidence of Lyme disease was reported (Table 6). The percent of high- and low-incidence counties correctly classified by the *B. burgdorferi* consensus map was 30.4% higher than the *I. scapularis* consensus map. The *B. burgdorferi* consensus map did not classify as many counties in the southeastern US as high-suitability compared to the *I. scapularis* map, yielding a lower false positive rate (Fig. 4).

4.7. Identifying leading edge counties

The *B. burgdorferi* consensus map more accurately aligns with high Lyme disease incidence counties than the *I. scapularis* map according to all three of our criteria: (1) better AUC and AIC values, (2) better discrimination in median incidence between high- and low-suitability counties, and (3) a higher percentage of high- and low-incidence counties correctly classified. Overall, when sensitivity is set to 90%, the *B. burgdorferi* consensus map identified 403 high-suitability counties for *B. burgdorferi*, but where Lyme disease incidence was not classified as high. These counties were primarily in the upper Midwest, specifically Ohio, Michigan, Indiana, Illinois, and Iowa. The median Lyme disease incidence in these counties was 0.87 cases per 100,000 persons (IQR: 0.36–2.48 cases per 100,000 persons) (Table 7).

To identify counties where *B. burgdorferi* may be detected with enhanced vector surveillance, or where enzootic transmission and incidence may increase over time, we set the sensitivity of the *B. burgdorferi* modeling algorithms to 95% (defining moderate-suitability). Median Lyme disease incidence in counties classified as moderate-suitability by the 95% sensitivity map was 3.37 cases per 100,000 persons (IQR: 0.66 – 23.60 cases per 100,000 persons). The difference in Lyme disease incidence was 3.27 cases per 100,000 persons between counties classified as moderately suitable and those classified as low-suitability. This map classifies an additional 186 leading edge counties as moderately suitable compared against the consensus map showing high-suitability set at 90% sensitivity. These counties were situated primarily in central portions of Iowa, Missouri, Kentucky and Tennessee, southern Virginia, northern North Carolina and coastal South Carolina (Fig. 4). The median Lyme disease incidence in these counties was 0.61 cases per 100,000 persons (IQR: 0.27–1.49 cases per 100,000 persons) (Table 7).

5. Discussion

We developed habitat suitability models for both *I. scapularis* and *B. burgdorferi*-infected host-seeking *I. scapularis*. The resulting consensus maps were used to identify counties where established populations of *I. scapularis* or presence of *B. burgdorfen*-infected *I. scapularis* might be under-reported and tick surveillance efforts should be directed. These counties were situated primarily in the upper Midwest (Illinois, Iowa, Indiana, Michigan, and Ohio). We confirmed that the distribution of counties classified as highly suitable for *B. burgdorferi*-infected *I. scapularis* more accurately predicts the distribution of high-incidence Lyme disease counties than the distribution of highly suitable habitat for the tick without accounting for infection status. When the sensitivity of the component models of the *B. burgdorferi* map were increased to 95%, we identified moderately suitable counties situated primarily in central portions of Iowa, Kentucky and Tennessee, central and northeastern portions of Missouri, southern Virginia, northern North Carolina and coastal South Carolina. Lyme disease incidence in these moderately suitable counties was slightly lower than that in high-suitability counties, but still higher than incidence in counties with low-suitability. This indicates that environmental conditions are likely suitable for *B. burgdorferi* to persist in ticks, but we currently lack records of *B. burgdorferi* in host-seeking ticks. This highlights the counties where enhanced tick surveillance may verify or refute the accuracy of model predictions. Although pathogen presence appears to be a reasonable predictor of high-incidence counties across the eastern US, more nuanced metrics, including tick infection prevalence and the density of host-seeking infected ticks may increase the accuracy of disease occurrence predictions, particularly in leading-edge counties.

Previous studies explored the predicted range of *I. scapularis* across the eastern US (Estrada-Peña, 2002; Brownstein et al., 2003; Hahn et al., 2016). Their predictions, and ours, demonstrate the dependence of model outcomes on the records upon which they are based, as well as the predictors used. Among previous modeling efforts, highly suitable habitat was consistently and accurately predicted in northeastern and southeastern states. Highly suitable tick habitat was also predicted along the Atlantic coast. The greatest uncertainty or disagreement among early suitability models was in the upper Midwest, primarily in Michigan, Indiana, Ohio and western and northern Minnesota where suitability was previously determined to be minimal (Estrada-Peña, 2002; Brownstein et al., 2003). Intensified surveillance and range expansion later showed *I. scapularis* was established in much of the upper Midwest, which was not initially considered suitable (Eisen et al., 2016). Our updated models and those of Hahn et al. (2016, 2017) that are based on updated surveillance records show a high degree of certainty in predicting highly suitable habitat for *I. scapularis* across the upper Midwest and expand the predicted range across Michigan, Ohio, Illinois, Iowa and Minnesota.

Surprisingly few studies have modeled the distribution of suitable habitat for the Lyme disease spirochetes. Data derived from the most comprehensive systematic tick surveillance effort in the US were used to estimate the density of host-seeking infected nymphs across the eastern US (Diuk-Wasser et al., 2012). However, in the ensuing 15 years since that study was completed, the range of both the tick and the pathogen have expanded (Eisen et al., 2016; Fleshman et al., 2021), necessitating an update. In addition, Diuk-Wasser

et al. (2012) focused primarily on modeling the density of host-seeking infected nymphs (DIN). Nymphs are arguably the most epidemiologically significant metric (Mather et al., 1996; Diuk-Wasser et al., 2012; Pepin et al., 2012), but focusing exclusively on this life stage likely under-estimates the presence of *B. burgdorferi* in ticks in the southeastern US where nymphs are seldom collected by drag sampling, but infected adults may bite humans (Stromdahl and Hickling, 2012; Diuk-Wasser et al., 2010; Arsnoe et al., 2015, 2019). Indeed, compared with earlier models focused on DIN (Diuk-Wasser et al., 2012) our models predict suitable habitat for *B. burgdorferi* infected ticks moderately further south. Our predictions also extend further north in Michigan, and further into western New York and Pennsylvania, likely due to the expansion of our dataset as more data have been collected across the eastern US.

Comparing our *I. scapularis* and *B. burgdorferi* consensus maps, both dichotomized at a fixed 10% omission rate, we show that the predicted range of infected ticks is a more accurate predictor of high-incidence Lyme disease counties. This is expected, as the presence of *B. burgdorferi*-infected *I. scapularis* is generally assumed to be a better predictor of human cases than the presence of *I. scapularis* alone (Eisen et al., 2016; Eisen and Paddock, 2021). Our analysis quantifies the accuracy of these acarological metrics for predicting high-incidence Lyme disease counties and identifies counties where we expect to find *B. burgdorferi*-infected host-seeking ticks with differing levels of certainty. Field efforts directed at counties most likely to be highly suitable, even with more conservative sensitivity cut-offs, are most likely to yield infected ticks, and should be prioritized for field collection efforts.

Determining the presence of *B. burgdorferi* in host-seeking *I. scapularis* is generally easier and less costly compared with measuring DIN, yet shows good alignment with counties that are classified as high-incidence for Lyme disease. Given the limited number of DIN records, we were not able to directly compare the accuracy of DIN versus pathogen presence alone in predicting the geographic distribution of Lyme disease incidence. A previous study (Pepin et al., 2012) showed that across the eastern US, modeled estimates of DIN were significantly and positively correlated with reported Lyme disease incidence, with DIN explaining approximately 70% of the variation in reported disease incidence. The association was strongest in states and counties with high prevalence of *B. burgdorferi* infection in host-seeking nymphs and the metric was most accurate at discriminating high- versus low-incidence counties; it yielded mixed results when predicting county-scale incidence within high-incidence states. Based on our models, *B. burgdorferi* presence appears to be a similarly accurate predictor of counties reporting high Lyme disease incidence. This is likely because at coarse spatial scales (county, state, or region) the prevalence of *B. burgdorferi* infection in *I. scapularis* tends to be relatively stable over time in areas where the spirochete has become established (Prusinski et al., 2014; Lehane et al., 2021; Foster et al., 2021). Specifically, nymphal and adult infection rates are commonly in the 20% and 50% range, respectively, in the Northeast, Mid-Atlantic and upper Midwest. By contrast, in the Southeast, host-seeking nymphs are seldomly collected and tested and adult infection prevalence is typically low, <5% (Diuk-Wasser et al., 2012; Lehane et al., 2021; Porter et al., 2021). Our dataset was based on counties where infection prevalence is relatively high and infected *I. scapularis* were more likely to be detected given limited

surveillance effort. Tick surveillance effort is not evenly distributed across states or counties in the eastern US. Therefore, the predicted suitable range of *B. burgdorferi* based on our models is likely an under-representation of the true range of the pathogen, particularly under-representing suitable habitat in areas where infection prevalence is low. Given our low positive predictive value (i.e., high incidence was predicted in 50% of counties classified as highly suitable for *B. burgdorferi*-infected ticks), additional effort is needed to assess if these counties will emerge as high-incidence, or if higher level acarological metrics (e.g., densities of infected host-seeking ticks) are needed to accurately predict incidence in these areas.

Similar to previous models, the predictors included in our *I. scapularis* models were focused on temperature, precipitation and vegetative cover. The included variables were similar to or highly correlated with those described in Hahn et al. (2016). Their biological plausibility is explained therein. Beyond the records included in Hahn et al. (2016), our dataset contained an additional 159 counties with established *I. scapularis* populations obtained through the recently initiated CDC tick surveillance program. This expanded dataset results in much of the Northeast, upper Midwest, and coastal Southeast being classified as suitable for *I. scapularis*, with the interior Southeast being classified as suitable at higher sensitivities (90% and 95%). All predictors used in the *I. scapularis* models were also used by the *B. burgdorferi* models, except elevation. Two additional predictors were also included in the *B. burgdorferi* models, the mean temperature of the wettest (BIO8) and driest (BIO9) quarters. It is difficult to assess why these variables were significant predictors of *B. burgdorferi* presence, but we speculate they either correlate with host distributions or variability in host-seeking phenology, each is assumed to affect prevalence in host-seeking ticks (Gatewood et al., 2009; Arsnoe et al., 2019; Ginsberg et al., 2021). Unfortunately, host distribution data are not available at the scale of our analyses. It is also worth noting that the scale of our analysis (i.e., county-level across the eastern US) likely affected the predictors that were selected, as interactions between ecosystem processes are partially dependent upon the scale at which data are organized (O'Neill et al., 1988, 1991; Newman et al., 2019). Hence, predictors used for finer scale modeling efforts may differ from those used by our models. Generally, hot and dry environments are not suitable for *I. scapularis* and the inclusion of these predictors in the *B. burgdorferi* models may be due to the fact that observations are clustered in the Northeast and upper Midwest, where precipitation is high, and temperatures are low. The exclusion of elevation as a predictor in the *B. burgdorferi* models may relate to there being observations of *B. burgdorferi*-infected *I. scapularis* in high and low elevation counties in the Northeast, as well as high elevation counties in West Virginia. This may have reduced the predictive power of elevation for *B. burgdorferi*-infected *I. scapularis*. Overall, fewer counties were classified as highly suitable for *B. burgdorferi*-infected *I. scapularis* than ticks when not accounting for infection status, and these were predominantly clustered in the upper midwestern and northeastern US.

Acarological risk data are expensive and time consuming to collect from the field, particularly at the national scale, but we have demonstrated that the collection of *B. burgdorferi* infection presence data for *I. scapularis* can accurately predict the geographic distribution of Lyme disease occurrence, particularly discriminating between low- and high-incidence counties. Because data were aggregated to the county scale, it is likely we are under-representing risk particularly in small suitable patches within counties along

the current leading edge. In some cases, it is also possible that entire counties may be classified as highly suitable, despite suitable habitat being limited to specific areas within the county. Additional analyses at sub-county scales are needed to determine the distribution of suitable habitat within counties classified as suitable. Furthermore, our analyses did not take spatial autocorrelation into account, hence it is not possible to determine whether our predictors are the underlying cause of the habitat being considered suitable, or this is an effect of a spreading population of *I. scapularis*. Models using spatial and habitat-based approaches should be compared, preferably using non-binary, sub-county data. To improve acarological risk estimates, tick surveillance efforts aimed at defining the distribution of *B. burgdorferi*-infected ticks should prioritize tick collection and testing in counties classified as suitable, but where infected ticks have not been reported. Counties classified as suitable at lower sensitivity thresholds (≤90%) are expected to be the highest yield for field efforts. Particularly in leading edge counties where *B. burgdorferi* has been detected but prevalence of infection has not been assessed, increased efforts to characterize infection rates may help to assess the likelihood that a tick-bite could result in transmission of *B. burgdorferi*. As national datasets continue to improve and grow, it is also important to determine what information is gained from the collection of more complex and costly metrics, particularly DIN, which accounts for both tick infection status and the likelihood of human-tick encounters.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank our numerous public health partners across the United States who worked to submit surveillance data to ArboNET or ticks to CDC for pathogen testing. The contents of this manuscript are solely the responsibility of the authors and do not necessarily represent the official views of CDC or the Department of Health and Human Services.

References

Aizpurua O, Cantú-Salazar L, San Martin G, Biver G, Brotons L, Titeux N, 2015. Reconciling expert judgement and habitat suitability models as tools for guiding sampling of threatened species. J. Appl. Ecol 52, 1608–1616.

Arsnoe IM, Hickling GJ, Ginsberg HS, McElreath R, Tsao JI, 2015. Different populations of blacklegged tick nymphs exhibit differences in questing behavior that have implications for human Lyme disease risk. PLoS ONE 10, e0127450. [PubMed: 25996603]

Arsnoe I, Tsao JI, Hickling GJ, 2019. Nymphal *Ixodes scapularis* questing behavior explains geographic variation in Lyme borreliosis risk in the eastern United States. Ticks Tick Borne Dis. 10, 553–563. [PubMed: 30709659]

Breiman L, 2001. Random forests. Mach. Learn 45, 5–32.

Brownstein JS, Holford TR, Fish D, 2003. A climate-based model predicts the spatial distribution of the Lyme disease vector *Ixodes scapularis* in the United States. Environ. Health Perspect 111, 1152–1157. [PubMed: 12842766]

Burnham KP, Anderson DR, 2004. Multimodel inference: understanding AIC and BIC in model selection. Sociol. Methods Res 33, 261–304.

Burtis JC, Sullivan P, Levi T, Oggenfuss K, Fahey TJ, Ostfeld RS, 2016. The impact of temperature and precipitation on blacklegged tick activity and Lyme disease incidence in endemic and emerging regions. Parasit. Vectors 9, 1–10. [PubMed: 26728523]

Burtis JC, Fahey TJ, Yavitt JB, 2019. Survival and energy use of *Ixodes scapularis* nymphs throughout their overwintering period. Parasitology 146, 781–790. [PubMed: 30638173]

CDC, 2018. Centers for Disease Control and Prevention Guidelines: "Surveillance for *Ixodes scapularis* and Pathogens Found in This Tick Species in the United States". CDC,. https://www.cdc.gov/ticks/resources/TickSurveillance_Iscapularis-P.pdf. Accessed 10 Aug 2021.

CDC, 2021a. Lyme Disease Surveillance and Available Data. Centers for Disease Control and Prevention, Atlanta, GA. https://www.cdc.gov/lyme/stats/survfaq.html. Accessed 04 Jan 2022.

CDC., 2021b. Lyme Disease (Borrelia Burgdorferi) Case Definitions. Centers for Disease Control and Prevention, Atlanta, GA. https://ndc.services.cdc.gov/conditions/lyme-disease/. Accessed 10 Aug 2021.

Crall AW, Jarnevich CS, Panke B, Young N, Renz M, Morisette J, 2013. Using habitat suitability models to target invasive plant species surveys. Ecol. Appl 23, 60–72. [PubMed: 23495636]

CSTE, 2007. Council of State and Territorial Epidemiologists. CSTE. Revised national surveillance case definition for Lyme disease. https://cdn.ymaws.com/www.cste.org/resource/resmgr/ps/ps2021/21-ID-05_Lyme_Disease.pdf. Accessed 10 Aug 2021.

Danielson JJ, Gesch DB, 2011. Global Multi-Resolution Terrain Elevation Data 2010 (GMTED2010) (p. 26). US Department of the Interior, US Geological Survey, Washington, DC, USA.

Dennis DT, Nekomoto TS, Victor JC, Paul WS, Piesman J, 1998. Reported distribution of *Ixodes scapularis* and *Ixodes pacificus* (Acari: ixodidae) in the United States. J. Med. Entomol 35, 629–638. [PubMed: 9775584]

Diuk-Wasser MA, Vourc'h G, Cislo P, Hoen AG, Melton F, Hamer SA, Rowland M, Cortinas R, Hickling GJ, Tsao JI, Barbour AG, Kitron U, Piesman J, Fish D, 2010. Field and climate-based model for predicting the density of host-seeking nymphal *Ixodes scapularis*, an important vector of tick-borne disease agents in the eastern United States. Glob. Ecol. Biogeogr 19, 504–514.

Diuk-Wasser MA, Hoen AG, Cislo P, Brinkerhoff R, Hamer SA, Rowland M, Cortinas R, Vourc'h G, Melton F, Hickling GJ, Tsao JI, Bunikis J, Barbour AG, Kitron U, Piesman J, Fish D, 2012. Human risk of infection with *Borrelia burgdorferi*, the Lyme disease agent, in eastern United States. Am. J. Trop. Med 86, 320–327.

Eisen RJ, Eisen L, Beard CB, 2016. County-scale distribution of *Ixodes scapularis* and *Ixodes pacificus* (Acari: ixodidae) in the continental United States. J. Med. Entomol 53, 349–386. [PubMed: 26783367]

Eisen L, Eisen RJ, 2016. Critical evaluation of the linkage between tick-based risk measures and the occurrence of Lyme disease cases. J. Med. Entomol 53, 1050–1062. [PubMed: 27330093]

Eisen RJ, Paddock CD, 2021. Tick and tickborne pathogen surveillance as a public health tool in the United States. J. Med. Entomol 58, 1490–1502. [PubMed: 32440679]

Elith J, Leathwick JR, Hastie T, 2008. A working guide to boosted regression trees. J. Anim. Ecol 77, 802–813. [PubMed: 18397250]

Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ, 2011. A statistical explanation of MaxEnt for ecologists. Divers. Distrib 17, 43–57.

Estrada-Peña A, 2002. Increasing habitat suitability in the United States for the tick that transmits Lyme disease: a remote sensing approach. Environ. Health Perspect 110, 635–640. [PubMed: 12117639]

Fielding AH, Bell JF, 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv 24, 38–49.

Fleshman AC, Graham CB, Maes SE, Foster E, Eisen RJ, 2021. Reported county-level distribution of Lyme disease spirochetes, *Borrelia burgdorferi* sensu stricto and *Borrelia mayonii* (Spirochaetales: spirochaetaceae), in host-seeking *Ixodes scapularis* and *Ixodes pacificus* Ticks (Acari: ixodidae) in the contiguous United States. J. Med. Entomol 58, 1219–1233. [PubMed: 33600574]

Fleshman AC, Foster E, Maes SE, Eisen RJ, 2022. Reported county-level distribution of seven human pathogens detected in host-seeking *Ixodes scapularis* and *Ixodes pacificus* (Acari: ixodidae) in the contiguous United States. J. Med. Entomol In press.

Forrester JD, Brett M, Matthias J, Stanek D, Springs CB, Marsden-Haug N, Oltean H, Baker JS, Kugeler KJ, Mead PS, Hinckley A, 2015. Epidemiology of Lyme disease in low-incidence states. Ticks Tick Borne Dis. 6, 721–723. [PubMed: 26103924]

Foster E, Burtis JC, Sidge JL, Tsao JI, Bjork J, Liu G, Neitzel DF, Lee X, Paskewitz S, Caporale D, Eisen RJ, 2021. Inter-annual variation in prevalence of *Borrelia burgdorferi* sensu stricto and *Anaplasma phagocytophilum* in host-seeking *Ixodes scapularis* (Acari: ixodidae) at long-term surveillance sites in the upper midwestern United States: implications for public health practice. Ticks Tick Borne Dis.. In press.

Fry J, Xian GZ, Jin S, Dewitz J, Homer CG, Yang L, Barnes CA, Herold ND, Wickham JD, 2011. Completion of the 2006 national land cover database for the conterminous United States. Photogramm. Eng. Remote Sens 77, 858–864.

Gatewood AG, Liebman KA, Vourc'h G, Bunikis J, Hamer SA, Cortinas R, Melton F, Cislo P, Kitron U, Tsao J, Barbour AG, Fish D, Diuk-Wasser MA, 2009. Climate and tick seasonality are predictors of *Borrelia burgdorferi* genotype distribution. Appl. Environ. Microbiol 75, 2476–2483. [PubMed: 19251900]

Ginsberg HS, Hickling GJ, Burke RL, Ogden NH, Beati L, LeBrun RA, Arsnoe IM, Gerhold R, Han S, Jackson K, Maestas L, Moody T, Pang G, Ross B, Rulison EL, Tsao JI, 2021. Why Lyme disease is common in the northern US, but rare in the south: the roles of host choice, host-seeking behavior, and tick density. PLoS Biol. 19, e3001066. [PubMed: 33507921]

Graham CB, Maes SE, Hojgaard A, Fleshman AC, Sheldon SW, Eisen RJ, 2018. A molecular algorithm to detect and differentiate human pathogens infecting *Ixodes scapularis* and *Ixodes pacificus* (Acari: ixodidae). Ticks Tick Borne Dis. 9, 390–403. [PubMed: 29258802]

Guisan A, Thuiller W, Zimmermann NE, 2017. Habitat Suitability and Distribution Models: With Applications in R. Cambridge University Press.

Hahn MB, Jarnevich CS, Monaghan AJ, Eisen RJ, 2016. Modeling the geographic distribution of *Ixodes scapularis* and *Ixodes pacificus* (Acari: ixodidae) in the contiguous United States. J. Med. Entomol 53, 1176–1191. [PubMed: 27282813]

Hahn MB, Jarnevich CS, Monaghan AJ, Eisen RJ, 2017. Response: the geographic distribution of *Ixodes scapularis* (Acari: ixodidae) revisited: the importance of assumptions about error balance. J. Med. Entomol 54, 1104–1106. [PubMed: 28874013]

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A, 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol 25, 1965–1978.

Johnson TL, Haque U, Monaghan AJ, Eisen L, Hahn MB, Hayden MH, Savage HM, McAllister J, Mutebi JP, Eisen RJ, 2017. Modeling the environmental suitability for *Aedes* (*Stegomyia*) *aegypti* and *Aedes* (*Stegomyia*) *albopictus* (Diptera: culicidae) in the contiguous United States. J. Med. Entomol 54, 1605–1614. [PubMed: 29029153]

Kugeler KJ, Farley GM, Forrester JD, Mead PS, 2015. Geographic distribution and expansion of human Lyme disease, United States. Emerg. Infect. Dis 21, 1455–1457. [PubMed: 26196670]

Lehane A, Maes SE, Graham CB, Jones E, Delorey M, Eisen RJ, 2021. Prevalence of single and coinfections of human pathogens in *Ixodes* ticks from five geographical regions in the United States, 2013–2019. Ticks Tick Borne Dis. 12, 101637. [PubMed: 33360805]

Linske MA, Stafford KC, Williams SC, Lubelczyk CB, Welch M, Henderson EF, 2019. Impacts of deciduous leaf litter and snow presence on nymphal *Ixodes scapularis* (Acari: ixodidae) overwintering survival in coastal New England, USA. Insects 10, 227. [PubMed: 31366124]

Liu C, White M, Newell G, 2013. Selecting thresholds for the prediction of species occurrence with presence-only data. J. Biogeogr 40, 778–789.

Mather TN, Nicholson MC, Donnelly EF, Matyas BT, 1996. Entomologic index for human risk of Lyme disease. Am. J. Epidemiol 144, 1066–1069. [PubMed: 8942438]

Merow C, Smith MJ, Edwards TC Jr, Guisan A, McMahon SM, Normand S, Thuiller W, Wüest RO, Zimmermann NE, Elith J, 2014. What do we gain from simplicity versus complexity in species distribution models? Ecography 37, 1267–1281.

Newman EA, Kennedy MC, Falk DA, McKenzie D, 2019. Scaling and complexity in landscape ecology. Front. Ecol. Evol 7, 293.

Ogden NH, Lindsay LR, Beauchamp G, Charron D, Maarouf A, O'callaghan CJ, Waltner-Toews D, Barker IK, 2004. Investigation of relationships between temperature and developmental rates of tick *Ixodes scapularis* (Acari: ixodidae) in the laboratory and field. J. Med. Entomol 41, 622–633. [PubMed: 15311453]

O'Neill RV, Milne BT, Turner MG, Gardner RH, 1988. Resource utilization scales and landscape pattern. Landsc. Ecol 2, 63–69.

O'Neill RV, Turner SJ, Cullinan VI, Coffin DP, Cook T, Conley W, Brunt J, Thomas JM, Conley MR, Gosz J, 1991. Multiple landscape scales: an intersite comparison. Landsc. Ecol 5, 137–144.

Pepin KM, Eisen RJ, Mead PS, Piesman J, Fish D, Hoen AG, Barbour AG, Hamer S, Diuk-Wasser MA, 2012. Geographic variation in the relationship between human Lyme disease incidence and density of infected host-seeking *Ixodes scapularis* nymphs in the Eastern United States. Am. J. Trop. Med 86, 1062.

Peterson AT, 2014. Mapping Disease Transmission Risk in Geographic and Ecological Contexts. Johns Hopkins University Press, Baltimore.

Peterson AT, Raghavan RK, 2017. The geographic distribution of *Ixodes scapularis* (Acari: ixodidae) revisited: the importance of assumptions about error balance. J. Med. Entomol 54, 1080–1084. [PubMed: 28591858]

Porter WT, Wachara J, Barrand ZA, Nieto NC, Salkeld DJ, 2021. Citizen science provides an efficient method for broad-scale Tick-Borne Pathogen Surveillance of *Ixodes pacificus* and *Ixodes scapularis* across the United States. mSphere 6, e00682–e00721. [PubMed: 34585963]

Prusinski MA, Kokas JE, Hukey KT, Kogut SJ, Lee J, Backenson PB, 2014. Prevalence of *Borrelia burgdorferi* (Spirochaetales: spirochaetaceae), *Anaplasma phagocytophilum* (Rickettsiales: anaplasmataceae), and *Babesia microti* (Piroplasmida: babesiidae) in *Ixodes scapularis* (Acari: ixodidae) collected from recreational lands in the Hudson Valley region, New York state. J. Med. Entomol 51, 226–236. [PubMed: 24605473]

R Core Team, 2021. R: A Language and Environment For Statistical Computing. R Foundation For Statistical Computing. R Core Team, Vienna, Austria. URL. https://www.R-project.org/.

Schneider SC, Parker CM, Miller JR, Page Fredericks L, Allan BF, 2015. Assessing the contribution of songbirds to the movement of ticks and *Borrelia burgdorferi* in the Midwestern United States during fall migration. EcoHealth 12, 164–173. [PubMed: 25297819]

Schwartz AM, Hinckley AF, Mead PS, Hook SA, Kugeler KJ, 2017. Surveillance for lyme disease—United States, 2008–2015. MMWR Surveill. Summ 66, 1–12.

Scott JD, 2016. Studies abound on how far north *Ixodes scapularis* ticks are transported by birds. Ticks Tick Borne Dis. 7 (2), 327–328. [PubMed: 26739029]

Springer YP, Jarnevich CS, Barnett DT, Monaghan AJ, Eisen RJ, 2015. Modeling the present and future geographic distribution of the lone star tick, *Amblyomma americanum* (Ixodida: ixodidae), in the continental United States. Am. J. Trop. Med 93, 875–890.

Stromdahl EY, Hickling GJ, 2012. Beyond Lyme: aetiology of tick-borne human diseases with emphasis on the South-Eastern United States. Zoonoses Public Healt 59, 48–64.

Talbert C, Talbert M, 2001. User Documentation For the Software for Assisted Habitat Modeling (SAHM) Package in VisTrails. US Geoglogical Survey, Fort Collins, CO.

Thornton PE, Running SW, White MA, 1997. Generating surfaces of daily meteorological variables over large regions of complex terrain. J. Hydrol 190, 214–251.

QGIS Development Team, 2021. QGIS geographic information system. open source geospatial foundation. URL http://qgis.org.

Thornton PE, Thornton MM, Mayer BW, Wei Y, Devarakonda R, Vose RS, Cook RB, 2016. Daymet: Daily Surface Weather Data On a 1-km Grid For North America, version 3.ORNL DAAC, Oak Ridge, Tennessee, USA.

US Census, 2021. Census Database. United States Census Bureau, US Census. New York, NY. http://www.census.gov/data.html. Accessed 15 Mar 2021.

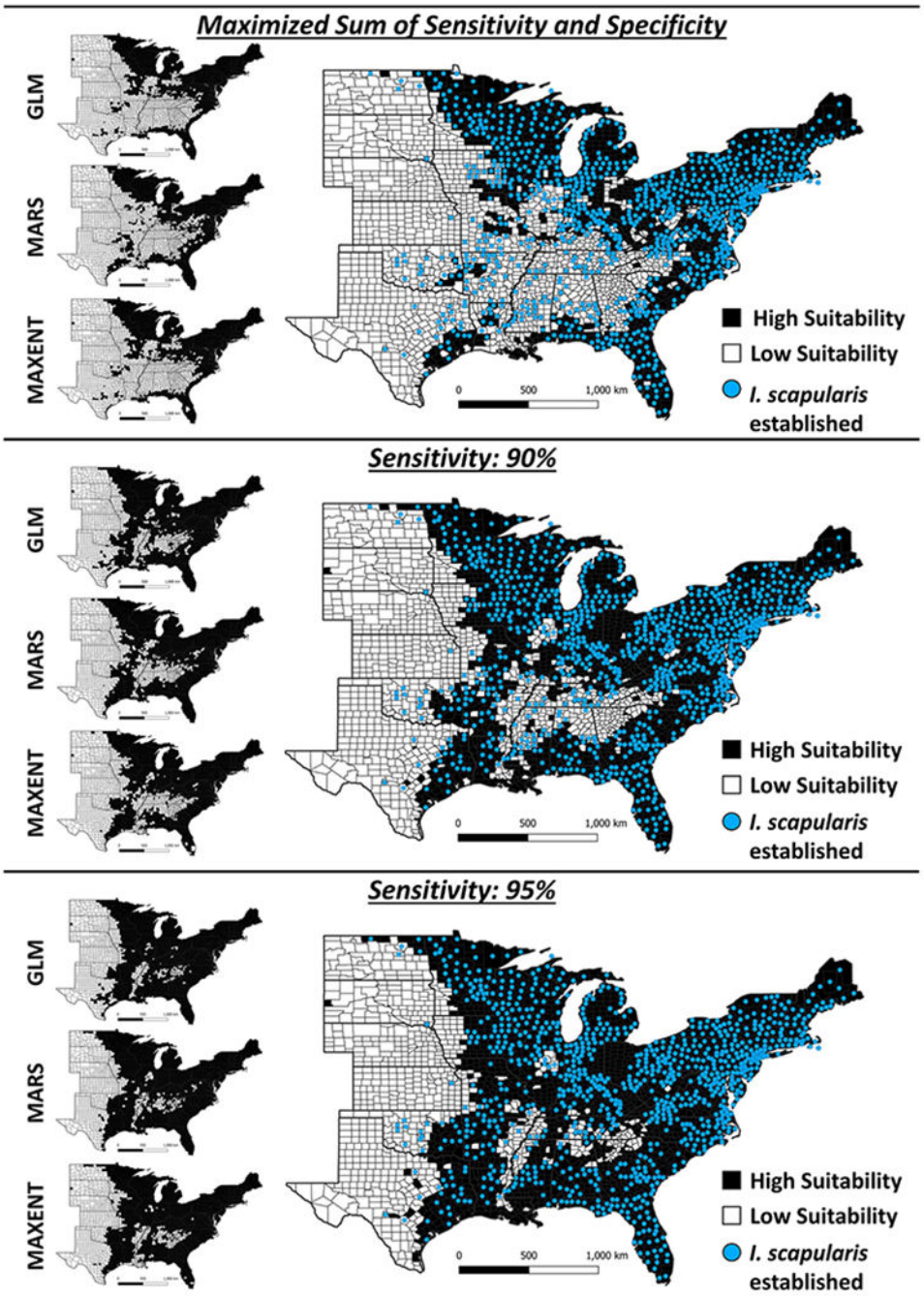

Fig. 1.

The predicted suitability of counties for *I. scapularis* establishment based on three individual models (GLM / MARS / MAXENT) shown on the left. The large consensus maps to the right show counties to be highly suitable when ≥ 2 of the individual models predict high-suitability. Colored points on the large maps represent counties where *I. scapularis* is established. The maps are shown at three levels of sensitivity: 76% (sum of sensitivity and specificity maximized), 90 and 95%.

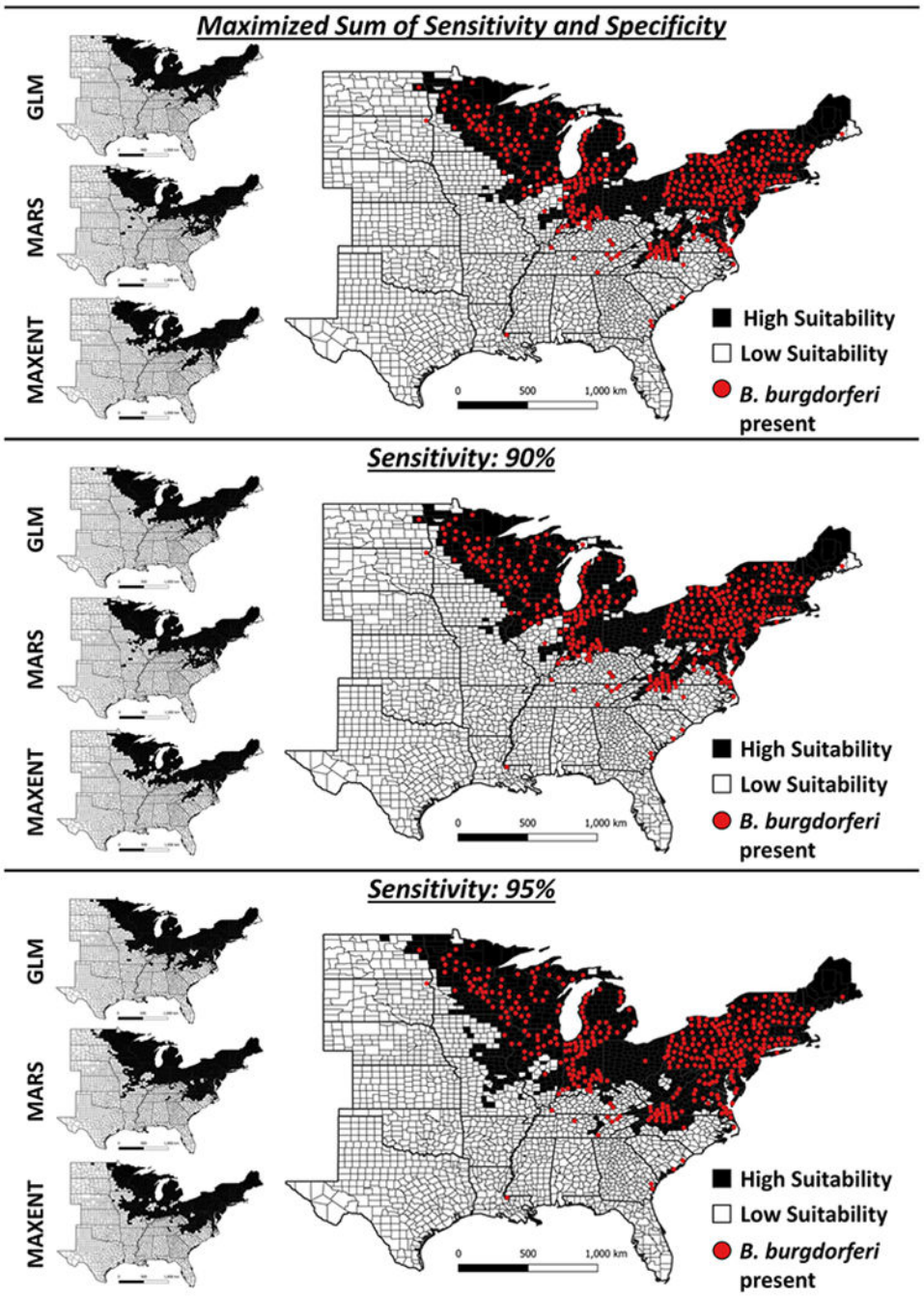

Fig. 2.

Counties predicted to be highly suitable for detecting *B. burgdorferi*-infected host-seeking *I. scapularis* based on three individual models (GLM/MARS / MAXENT), shown in maps on the left. The large consensus maps to the right show counties to be highly suitable when ≥2 of the individual models predict high-suitability. Colored points on the large maps represent counties where *B. burgdorferi* was detected in field collected host-seeking *I. scapularis*. The maps are shown at three levels of sensitivity: 88% (sum of sensitivity and specificity maximized), 90, and 95%.

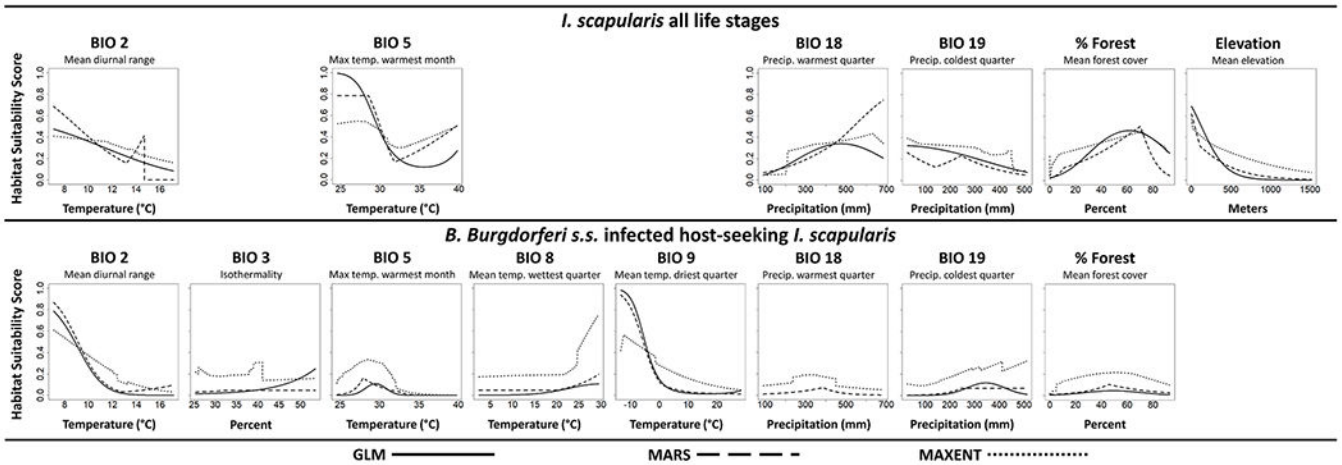

Fig. 3.

Response curves for the predictive variables included in the climate suitability models using the two data sets; *I. scapularis* and *B. burgdorferi*. The different line types represent the modeling algorithms, solid lines are GLM, dashed lines are MARS, and dotted lines are MAXENT. Not all parameters were used in all models.

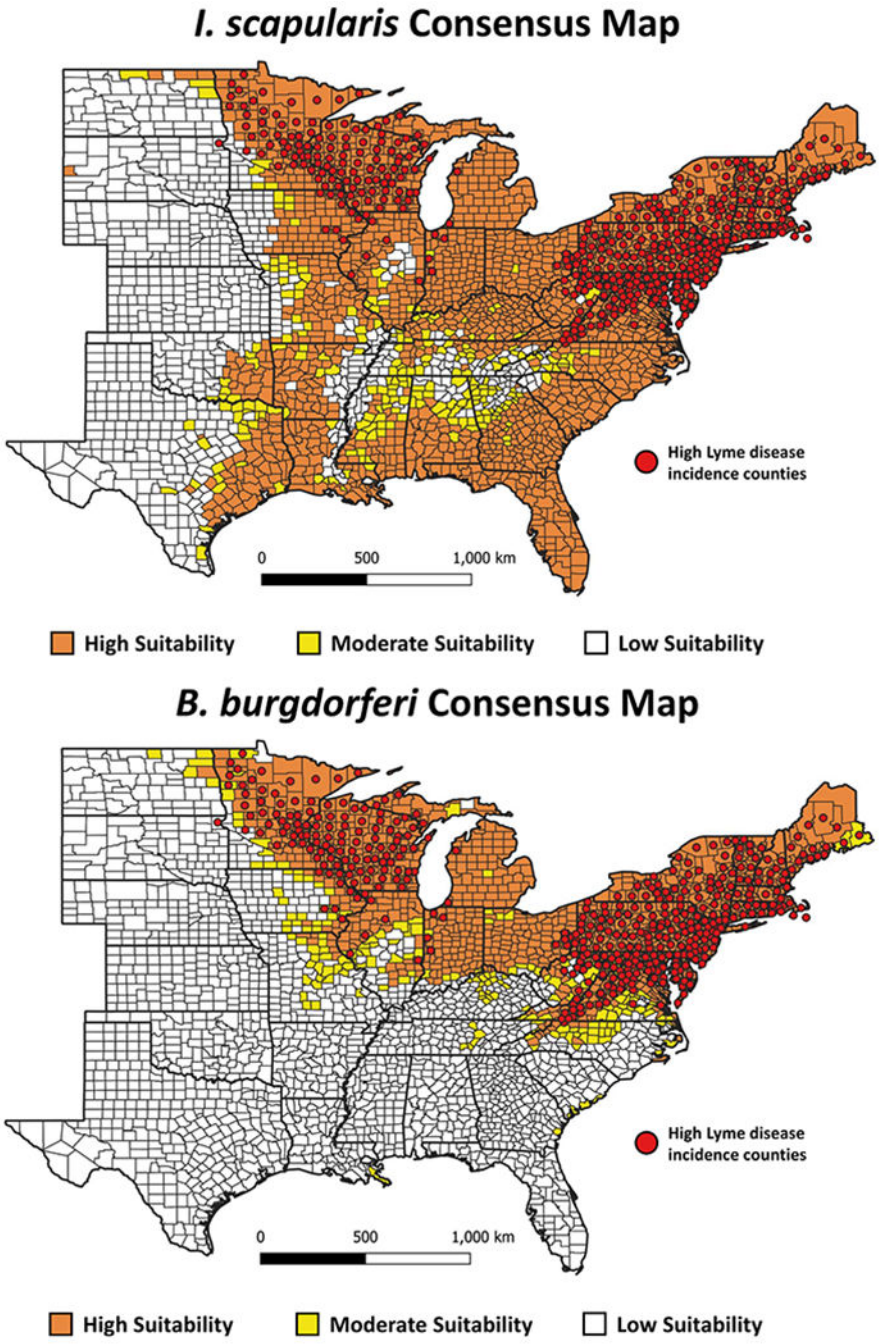

Fig. 4.

Counties predicted to be suitable by the *I. scapularis* (top) and *B. burgdorferi* (bottom) consensus maps. High-suitability counties are based on models set to 90% sensitivity (orange), moderate-suitability counties with models set to 95% sensitivity (yellow). Low-suitability counties are shown in white. Red dots indicate counties reporting high incidence of Lyme disease (≥10 cases per 100,000 persons) for at least three consecutive years

between 2000 and 2019 (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

Table 1

The 24 climate and landscape predictors included in the initial parameter selection for models based upon both field datasets. The ‘mean min max’ column shows whether predictor maximums, minimums, or means were averaged at the county level.

Variable Name	Min Mean Max	Brief Description	Data Source
Bio1 (°C)	Mean	Annual mean temperature	Daymet
Bio2 (°C)	Mean	Mean diurnal range	Daymet
Bio3 (%)	Max	Isothermality (BIO2/BIO7) (x100)	Daymet
Bio4 (°C)	Max	Temperature seasonality (Std Dev x 100)	Daymet
Bio5 (°C)	Max	Maximum temperature of the warmest month	Daymet
Bio6 (°C)	Min	Minimum temperature of coldest month	Daymet
Bio7 (°C)	Max	Annual range of temperature	Daymet
Bio8 (°C)	Mean	Mean temperature of wettest quarter	Daymet
Bio9 (°C)	Mean	Mean temperature of driest quarter	Daymet
Bio10 (°C)	Max	Maximum temperature of warmest quarter	Daymet
Bio11 (°C)	Min	Minimum temperature of coldest quarter	Daymet
Bio12 (°C)	Mean	Annual precipitation	Daymet
Bio13 (mm)	Max	Precipitation of wettest month	Daymet
Bio14 (mm)	Min	Precipitation of driest month	Daymet
Bio15 (mm)	Max	Precipitation seasonality (coefficient of variation)	Daymet
Bio16 (mm)	Max	Precipitation during wettest quarter	Daymet
Bio17 (mm)	Min	Precipitation during driest quarter	Daymet
Bio18 (mm)	Mean	Precipitation during warmest quarter	Daymet
Bio19 (mm)	Mean	Precipitation during coldest quarter	Daymet
GDD10 DEC-FEB (days)	Mean	Average growing degree days at 10 °C from DEC - FEB	Daymet
Monthly SWE NOV-APR (mm)	Max	Snow water equivalent between NOV - APR	Daymet
Monthly VP MAR-JUN (kPa)	Min	Vapor pressure between MAR - JUN	Daymet
Percent forest (%)	NA	Percent area covered by forest	USGS
DEM (m)	Mean	Average elevation	USGS

*The spatial resolution of all data layers was 1 km × 1 km.

**All Daymet climate variables were derived based on data from 1980 to 2015, forest cover was based on 2011 data and elevation was based on data collected in 2006.

Table 2

Model selection criteria and performance metrics for the testing and training runs of each modeling algorithm used to construct the consensus models based upon the two acarological risk metrics; *I. scapularis* and *B. burgdorferi*-infected *I. scapularis*.

***I. scapularis* suitability models**						
Performance metric	**GLM**		**MARS**		**Maxent**	
	Test	**Train**	**Test**	**Train**	**Test**	**Train**
AUC	0.84	0.84	0.84	0.84	0.84	0.86
Percent correctly classified	75.3	75.4	75.2	75.5	75.6	76.1
Mean threshold	0.41	0.39	0.40	0.37	0.43	0.41
Sensitivity	68.9	73.7	71.5	75.5	72.7	78.5
Specificity	79.1	76.4	77.3	75.8	77.3	74.7
PPV	66.1	64.8	65.1	64.7	65.4	64.7
NPV	81.2	83.1	82.1	83.8	82.8	85.5
Correlation coefficient	0.57	0.58	0.57	0.59	0.55	0.57
Percent deviance explained	27.9	38.1	28.5	30.0	23.8	25.5
***B. burgdorferi* suitability models**						
Performance metric	**GLM**		**MARS**		**Maxent**	
	Test	**Train**	**Test**	**Train**	**Test**	**Train**
AUC	0.92	0.92	0.91	0.92	0.92	0.94
Percent correctly classified	83.4	80.4	78.7	81.4	83.0	84.5
Mean threshold	0.17	0.11	0.11	0.14	0.32	0.33
Sensitivity	86.0	92.5	90.0	89.8	85.1	89.1
Specificity	82.9	78.2	76.7	80.0	82.6	83.7
PPV	46.9	42.7	40.4	44.0	46.2	48.9
NPV	97.1	98.4	97.8	97.8	96.9	97.8
Correlation coefficient	0.62	0.63	0.61	0.63	0.60	0.64
Percent deviance explained	41.6	43.7	39.8	42.6	37.2	41.4

AUC: *The AUC is a measure of model accuracy. A value of 1 indicates a 'perfect' model and values ≤ 0.5 indicate a poor distinction between counties classified as high- or low-suitability.*

Percent Correctly Classified: *(True Positive + True Negative) / (True Positive + True Negative + False Positive + False Negative).*

Mean threshold: *Probability threshold at which presence is with the sum of sensitivity and specificity maximized.*

Sensitivity: *True Positive / (True Positive + False Negative).*

Specificity: *True Negative / (True Negative + False Positive).*

Positive Predictive Value (PPV) = *True Positive / (True Positive / False Positive).*

Negative Predictive Value (NPV) = *True Negative / (False Negative + True Negative).*

Correlation Coefficient: *Linear relationship between the field data and model output.*

Percent Deviance Explained: *Goodness of fit statistic, similar to an R^2 value.*

Table 3

Relative contributions of the climate predictors selected by the distribution modeling algorithms for the two acarological risk metrics; *I. scapularis* and *B. burgdorferi*-infected *I. scapularis*.

***I. scapularis* suitability models**				
Predictors	**Percent Deviance Explained**	**Normalized contribution values (%)**		
		GLM	**MARS**	**Maxent**
Mean diurnal temp. range (BIO2)	12.0	1.2	12.4	4.3
Max temp. warmest month (BIO5)	9.1	27.5	29.9	23.1
Precip. of coldest quarter (BIO19)	8.3	4.1	4.6	7.6
Precip. of warmest quarter (BIO18)	7.5	10.5	6.9	13.7
Percent forest cover	7.2	30.6	30.5	33.5
Elevation	5.3	26.1	15.7	17.8
***B. burgdorferi* suitability models**				
Predictors	**Percent Deviance Explained**	**Normalized contribution values (%)**		
		GLM	**MARS**	**Maxent**
Max temp. warmest month (BIO5)	33.1	32.5	23.4	31.3
Mean diurnal temp. range (BIO2)	22.7	11.5	8.3	7.3
Isothermality (BIO3)	17.7	1.5	0.2	2.1
Precip. of coldest quarter (BIO19)	13.1	14.2	16.4	6.3
Precip. of warmest quarter (BIO18)	10.5	–	2.2	2.1
Mean temp. of driest quarter (BIO9)	18.9	27.4	37.8	24.8
Percent forest cover	7.3	8.9	10.6	22.3
Mean temp. of wettest quarter (BIO8)	5.7	4.0	1.1	4.0

Table 4

Sensitivity, specificity, positive predictive values (PPV), negative predictive value (NPV), and percent correctly classified (PCC) for the two consensus maps against the *I. scapularis* and *B. burgdorferi* binary (established / not established) datasets, with the sensitivities of their three component models set at three different levels. The model settings refer to the sensitivities set for the GLM, MARS, and Maxent models.

Consensus Layers	Model Settings	Sensitivity	Specificity	PPV	NPV	PCC
***I. scapularis* consensus map**	Sens+spec	76.9%	75.1%	64.7%	84.6%	75.8%
	Sens: 90%	90.5%	54.0%	53.7%	90.6%	67.5%
	Sens: 95%	95.4%	42.7%	49.6%	94.0%	62.3%
***B. burgdorferi* consensus map**	Sens+spec	91.0%	81.0%	45.7%	98.1%	82.5%
	Sens: 90%	89.8%	80.7%	45.0%	97.8%	82.1%
	Sens: 95%	95.2%	72.3%	37.6%	98.9%	75.7%

*Sens+spec is the output which maximizes the sum of sensitivity and specificity.

Table 5

Median, 25%, and 75% quartiles of human Lyme disease incidence in counties classified as high- and low-suitability by the consensus maps based upon the two acarological risk metrics; *I. scapularis* and *B. burgdorferi*-infected *I. scapularis*. The median and quartile incidences of counties where these consensus maps overlap are also shown. Consensus map coverage shown here with the sensitivity of the models set to 90%.

Consensus Layers	Suitability	Median LD Incidence	25% Quartile	75% Quartile	Number Of Counties
***I. scapularis* consensus map**	High-Suitability	0.672	0.154	6.204	1686
	Low-Suitability	0.062	0.000	0.391	1009
***B. burgdorferi* consensus map**	High-Suitability	5.706	0.867	30.852	803
	Low-Suitability	0.149	0.000	0.485	1892
Regional Median	-	***0.338***	***0.000***	***1.644***	***2695***

Table 6

Percent of high and low Lyme disease incidence counties correctly classified by the consensus maps predicting presence with the sensitivity of the models set to 90%. High-incidence counties are those with at least three consecutive years between 2000 and 2019 with ≥ 10 confirmed Lyme disease cases per 100,000 persons.

***I. scapularis* consensus map**			
Map classification	**Incidence classification**		
	High Incidence	***Low Incidence***	***% Correct***
High-Suitability	433	1253	***25.68 (PPV)***
Low-Suitability	3	1006	***99.70 (NPV)***
% ***Correct***	***99.31 (sensitivity)***	***44.53 (specificity)***	***53.40 (PCC)***
***B. burgdorferi* consensus map**			
Map classification	**Incidence classification**		***% Correct***
	High Incidence	**Low Incidence**	
High-Suitability	403	400	***50.19 (PPV)***
Low-Suitability	36	1856	***98.10 (NPV)***
% ***Correct***	***91.80 (sensitivity)***	***82.27 (specificity)***	***83.81 (PCC)***

Table 7

The Lyme disease incidence within high-incidence Lyme disease counties, counties predicted to be highly suitable by the *B. burgdorferi*-infected host-seeking *I. scapularis* consensus map at 90% sensitivity, moderately suitable counties at 95% sensitivity, and low-suitability counties.

	Median LD Incidence	*25% Quartile*	*75% Quartile*	*Number of Counties*
High-Incidence Counties	29.59	14.01	55.73	436
High-Suitability	0.868	0.362	2.479	403
Moderate-Suitability	0.612	0.273	1.486	186
Low-Suitability	0.104	0.000	0.386	1670