Peer Community Journal

Section: Ecology

RESEARCH ARTICLE

Published
2023-12-06

Cite as
William Manley, Tam Tran, Melissa Prusinski and Dustin Brisson (2023) *Modeling Tick Populations: An Ecological Test Case for Gradient Boosted Trees*, Peer Community Journal, 3: e116.

Correspondence
wmanl@sas.upenn.edu

Peer-review
Peer reviewed and recommended by PCI Ecology, https://doi.org/10.24072/pci.ecology.100532

Modeling Tick Populations: An Ecological Test Case for Gradient Boosted Trees

William Manley [1], Tam Tran [1], Melissa Prusinski [2], and Dustin Brisson [1]

Volume **3** (2023), article e116

https://doi.org/10.24072/pcjournal.353

Abstract

General linear models have been the foundational statistical framework used to discover the ecological processes that explain the distribution and abundance of natural populations. Analyses of the rapidly expanding cache of environmental and ecological data, however, require advanced statistical methods to contend with complexities inherent to extremely large natural data sets. Modern machine learning frameworks such as gradient boosted trees efficiently identify complex ecological relationships in massive data sets, which are expected to result in accurate predictions of the distribution and abundance of organisms in nature. However, rigorous assessments of the theoretical advantages of these methodologies on natural data sets are rare. Here we compare the abilities of gradient boosted and linear models to identify environmental features that explain observed variations in the distribution and abundance of blacklegged tick (*Ixodes scapularis*) populations in a data set collected across New York State over a ten-year period. The gradient boosted and linear models use similar environmental features to explain tick demography, although the gradient boosted models found non-linear relationships and interactions that are difficult to anticipate and often impractical to identify with a linear modeling framework. Further, the gradient boosted models predicted the distribution and abundance of ticks in years and areas beyond the training data with much greater accuracy than their linear model counterparts. The flexible gradient boosting framework also permitted additional model types that provide practical advantages for tick surveillance and public health. The results highlight the potential of gradient boosted models to discover novel ecological phenomena affecting pathogen demography and as a powerful public health tool to mitigate disease risks.

[1]Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania, USA, [2]New York State Department of Health, Albany, New York, USA

Peer Community Journal is a member of the Centre Mersenne for Open Scientific Publishing
http://www.centre-mersenne.org/

e-ISSN 2804-3871

Contents

Introduction

Statistical models have been a cornerstone of understanding ecological phenomena in the natural world. Ecological models traditionally focus on identifying the biotic and abiotic drivers of natural phenomena and on explaining the distribution and abundance of populations (Austin et al., 1984; Elith and Leathwick, 2009; Harvey et al., 1980; McLain et al., 1995; Tran et al., 2021a). Classical generalized linear modeling has resulted in many foundational ecological discoveries (Abbott et al., 1977; Austin et al., 1990; Elith and Leathwick, 2009; Kleiber, 1947; Root, 1988; Tilman et al., 1996). This modeling framework, however, has several technical disadvantages including strict assumptions about error distributions, sensitivity to outliers, and an assumption of linear relationships between variables that can limit predictive power (Hastie et al., 2001; McCullagh and Nelder, 1989; Naghibi and Pourghasemi, 2015; Olden et al., 2008; Yee and Mitchell, 1991). The introduction of machine learning methods such as gradient boosted trees overcomes many of these limitations, although direct comparisons of the effectiveness of machine learning methods and linear models on natural data sets are rare (De'ath, 2007; Elith et al., 2008; Elith et al., 2006; Friedman, 2001). In this study, we compare a gradient boosting machine learning method (Pedregosa et al., 2011) with comparable general linear models in their ability to identify environmental features affecting population dynamics and their ability to predict the distribution and abundance of blacklegged ticks (*Ixodes scapularis*), an arthropod vector of multiple human pathogens.

Many machine-learning frameworks such as neural networks, random forests, and gradient boosted trees are well suited to investigate ecological phenomena in the increasingly data-rich research environment (Cutler et al., 2007; Farley et al., 2018; Friedman, 2001; Han et al., 2015; Rammer and Seidl, 2019; Stephens et al., 2017; Tran et al., 2021b). Among machine learning methods gradient boosted trees are well reputed for very high predictive accuracy and accurate identification of nonlinear relationships on tabular data (Bentéjac et al., 2021; Elith et al., 2008; Grinsztajn et al., 2022). Gradient boosting is an efficient machine learning algorithm that can analyze large data sets, identify complex relationships among variables, and make highly accurate spatio-temporal forecasts. The power of the gradient boosting algorithm is in part derived from their ability to automatically identify non-linear and non-additive relationships by combining hundreds of decision trees into a highly accurate ensemble (De'ath, 2007; De'ath and Fabricius, 2000). These models have several advantages over traditional linear models including that they accept many data types, are unconstrained by data and error distributions, and automatically detect nonlinear and interactive relationships. Further, cross-validation and advances in interpretative machine learning algorithms have addressed prior concerns that gradient boosted algorithms are prone to over-fitting and are too complex to derive ecological inferences (Elith et al., 2008; Lundberg and Lee, 2017; Rudin, 2019; Ryo et al., 2021).

The ability of linear and gradient boosted models to identify ecologically relevant features or to forecast demographic changes is rarely assessed in natural systems, despite the availability of appropriate data sets (though see (Becker et al., 2020; Elith et al., 2006; Escobar et al., 2018; Qiao et al., 2015; Shabani et al., 2016)). On one such dataset, linear models that explored the explanatory power of 217 environmental variables on the distribution and abundance of *I. scapularis* ticks identified several geographical, temporal, seasonal, environmental, climatic, and landscape features that accounted for the majority of the natural variance in tick demography (Tran et al., 2021a). These linear models accurately predicted the distribution and abundance of tick populations in future years, providing a potentially powerful public health tool to mitigate human disease risks from *I. scapularis*-borne pathogens including the agents causing Lyme disease, babesiosis, and anaplasmosis (Burgdorfer et al., 1982; Spielman et al., 1979; Telford et al., 1996). However, the data distributions assumed in this linear model framework required separate distribution and abundance models and the default assumptions of linearity and additivity limited the exploration of non-linear and non-additive effects which are ubiquitous in ecological systems (Hastie et al., 2001; Levin, 1998; McCullagh and Nelder, 1989; Olden et al., 2008; Tran et al., 2021a; Yee and Mitchell, 1991).

Here, we use gradient boosted trees to investigate the relationship between environmental features and the distribution and abundance of *I. scapularis* using the same dataset previously analyzed with general linear models (Tran et al., 2021a). The gradient boosted models were used to forecast the distribution and abundance of ticks in areas and years not used to build the models. Both the environmental features determined to influence tick demographics and the predictive performance of the gradient boosted tree models were compared to linear models trained and validated using the same data sets (Tran et al., 2021a). Additionally, we utilize the flexibility of the gradient boosting framework to build and validate two additional models that offer practical benefits for disease surveillance, including ease of interpretation and the ability to simultaneously predict tick distribution and abundance.

Methods

Study System

The presence and abundance of host-seeking nymphs were determined at 532 unique locations between 2008 and 2018 using the standardized dragging, flagging, and walking survey protocols described previously (Prusinski et al., 2014; Tran et al., 2021a). Locations were sampled every 1–5 years with an average of 4.7 visits per site between 2008 and 2018. The environmental features investigated as explanatory factors in our statistical models can be broadly categorized as geographical, temporal, seasonal, climatic, and landscape features. The tick density and environmental data used in this study are identical to those previously described in Tran et al. (2021a) to rigorously evaluate the relative efficacy of the gradient boosted and linear statistical models.

Distribution and Abundance Models

Independent distribution and abundance gradient boosted models were built to allow direct comparisons with the previously published distribution and abundance linear models (Tran et al., 2021a). A combined distribution and abundance linear model was not built, as a log-transformation of tick abundance was used to approximate a normal distribution and thus sites where ticks were absent could not be accommodated (Tran et al., 2021a). Data were also processed as described previously in Tran et al. (2021a) to aid comparisons between gradient boosted and linear models. As examples, ticks were considered "present" at a site in a given year if nymphs were detected at any of the multiple site visits within the year and the visit with the greatest nymphal abundance estimate was used as the abundance value for that site in that year. For a summary of built models see (Supplemental Table 2).

Training of gradient boosted models included feature selection, hyper-parameter tuning, and model fitting to the training data set (data from 2008-2017). Environmental features were selected separately for each model using a step-forward feature selection algorithm that optimizes

average predictive performance on a 5-fold cross-validation data set (Raschka, 2018). Briefly, each of the 5 folds of the cross-validation data set was generated by randomly partitioning the training data into subsets for model fitting (80% of data) and evaluation (20% of data), such that each fold would contain a unique 20% of the training data for evaluation. Models were limited to 30 or fewer environmental features to reduce the probability of over-fitting (Cawley and Talbot, 2010). Hyper-parameters that influence the learning process were tuned using a random search algorithm to find values that maximized performance on cross-validation data sets (Pedregosa et al., 2011). Using cross-validation sets to optimize which features and hyper-parameters are used in the final model fitting process reduces over-fitting to the training data, making the resultant model more likely to generalize to out-of-sample data (data collected in 2018, which was not used to train the model). The analytical code for this training process is available at (https://doi.org/10.17632/w8bp678m3f.2; Manley et al., 2023).

Predictive Accuracy Assessment

The out-of-sample predictive accuracy of the gradient boosted distribution and abundance models was compared to the accuracy of linear distribution and the abundance models using the previously published accuracy metrics (Tran et al., 2021a). Briefly, the predictions from gradient boosted and linear distribution models to the 2018 out-of-sample data were assessed based on accuracy, sensitivity, and specificity. Abundance model predictions to the out-of-sample data were compared using root-mean-squared-error and R^2 values. Additionally, to compare the abundance models in accordance with the methodology from Tran et al. (2021a), abundances were converted from log-transformed counts of nymphs into discrete categories of low (1-4 nymphs), medium (7-35), and high (36+), and predictions were considered accurate if they were within one natural log unit of the average prediction error.

Simultaneous Modeling of Distribution and Abundance

A multi-class categorical model and a density-estimating regression model were built using the gradient boosting framework. These models do not require the data processing, such as the log-transformation necessary for the linear models, which allows simultaneous analysis of presence and abundance from all sites and years. The multi-class model predicts nymphal abundance to one of three categories: absent (no nymphs), low abundance (1-35 nymphs), and high abundance ($>$35 nymphs). Out-of-sample performance was assessed as the accuracy of the predicted classification to locations visited in 2018.

The gradient boosted density model is similar to the previously described abundance model except that the response variable was tick density, as opposed to the number of ticks collected used in the linear model, and that site densities of zero ticks were permitted. Nymphal density was estimated as the number of ticks collected per collection-hour. Collector hours here were limited to four as preliminary analyses and prior studies demonstrated that density estimates were biased when larger collection-hour values were included (Tran et al., 2021a). The statistical weight of sites during model fitting was positively correlated with collection-hour up to four hours as density estimate accuracy is greater at sites with more sampling effort.

Environmental Feature Analyses

The relationships between nymphal tick distribution or abundance with individual environmental features in each model were analyzed using SHAP (SHapley Additive exPlanation) values (Lundberg and Lee, 2017). Briefly, this interpretative framework estimates the impact each model feature has on model predictions. Together these estimates provide a global view of the impact of each feature on model predictions in the context of other model features. SHAP values were used to identify and visualize the non-linear relationships and interaction effects discovered by each model. SHAP values were not used to evaluate the impact of environmental variables on predictions from the multi-class model as the complex outputs of this model are not supported in this analytical framework.

Results

The gradient boosted distribution and abundance models outperformed their linear model counterparts in both predictive power and identification of complex relationships between environmental features. The gradient boosted distribution model (Figure 1A), built using data from 2008-2017, accurately predicted 94% of sites where ticks were present in 2018 and 84% of sites where ticks were absent. By comparison, the linear distribution model trained and tested on the same data accurately predicted 80.6% of sites where ticks were present and 80.7% of sites where they were absent. Importantly, the gradient boosted model had a far lower false negative rate than the linear model (5.8% vs 19.4%), an especially costly error for public health efforts. The gradient boosted distribution model also made highly accurate predictions to the 27 sites that were visited for the first time in 2018 (true positive rate = 85%; true negative rate = 86%).

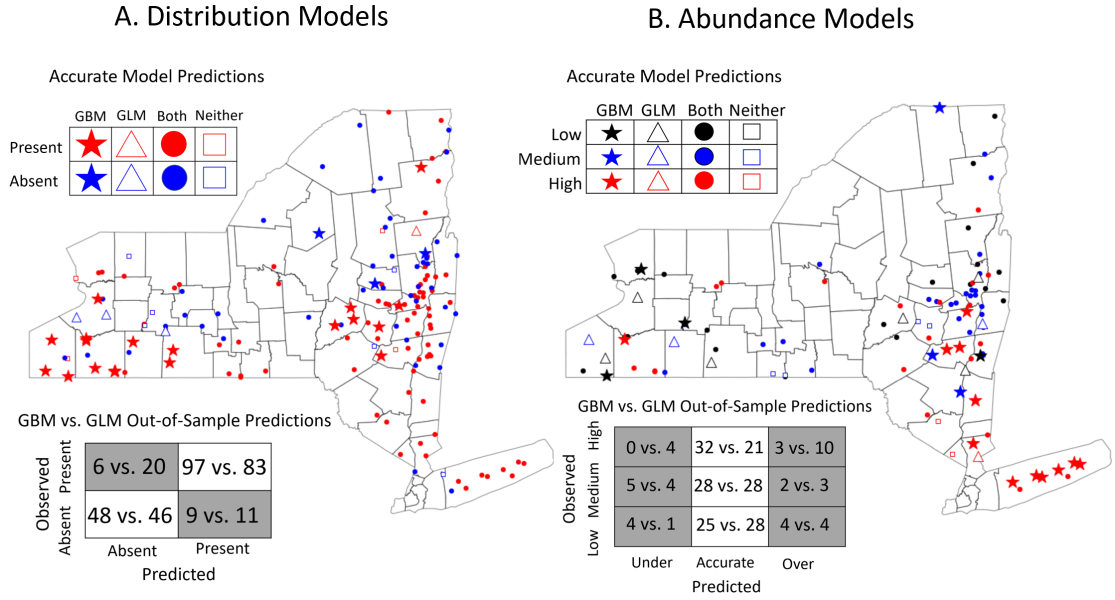

Figure 1 – Gradient boosted models more accurately predict future (A) distributions and (B) abundances of nymphal ticks than generalized linear models. (A) The gradient boosted distribution model was more accurate (90.6% vs 80.6%), more sensitive (true positive rate = 94.2% vs 80.5%), and more specific (true negative rate = 84.2% vs 80.7%) than its linear model analog. (B) The gradient boosted abundance model also more accurately predicted to the out-of-sample data than its linear model counterpart (82.5% vs 74.8%). Stars indicate sites with accurate predictions from the gradient boosted model and inaccurate predictions from the linear model; triangles represent accurate linear model predictions and inaccurate gradient boosted model predictions; squares represent sites accurately predicted by both models; circles represent inaccurate predictions by both models. Confusion matrices summarize the accurate and inaccurate predictions made by the gradient boosted model vs the linear model.

The gradient boosted abundance model more accurately predicted out-of-sample tick abundance than the analogous linear model in all quantitative metrics (RMSE = 0.972 vs. 1.096; R^2 = 0.59 vs. 0.48). Gradient boosted model predictions were also converted into discrete categories to compare the accuracy of the linear and gradient boosted models using the previously published methodology (Tran et al., 2021a). The gradient boosted abundance model was more accurate than its linear model counterpart, correctly predicting the abundance at 82.5% of sites compared to the 74.8% of sites correctly predicted by the linear model (Figure 1B). Sites visited for the first time in 2018 were also predicted with high accuracy by the gradient boosted model (83.3%; RMSE = 0.948; R^2 = 0.61). Importantly, nearly 40% of all sites incorrectly predicted by the gradient boosted model were conservative in that the model overestimated tick abundances at sites with high abundance (n=3) or underestimated tick abundance at sites with

low abundance (n=4). These errors are less costly as they indicate that the model has correctly predicted sites with high or low tick abundance but erred in terms of magnitude.

Complex non-linear relationships between environmental features and nymphal abundance were detected in gradient boosted models that were not investigated in the previously published linear models (Tran et al., 2021a). For example, estimates of deer population size have a highly complex relationship with nymphal abundance (Figure 2A): deer harvest values less than 2000 result in decreased nymphal abundance predictions; deer harvest between 2000 and 3000 are correlated with increases in nymphal abundances; deer harvest between 3000 and 6000 are correlated with decreased nymphal abundances; and deer harvest above 6000 is correlated with increased nymphal abundance. Although not biologically relevant, the number of tick collection efforts (sampling hours) had a positive but decelerating relationship with the number of nymphs collected (Figure 2B). That is, the number of nymphs collected is strongly and positively correlated with the number of hours field technicians flagged for ticks at sites visited for fewer than two hours. However, this positive relationship becomes less pronounced at sites visited for greater than two hours and is not detectable at sites visited for more than five hours.

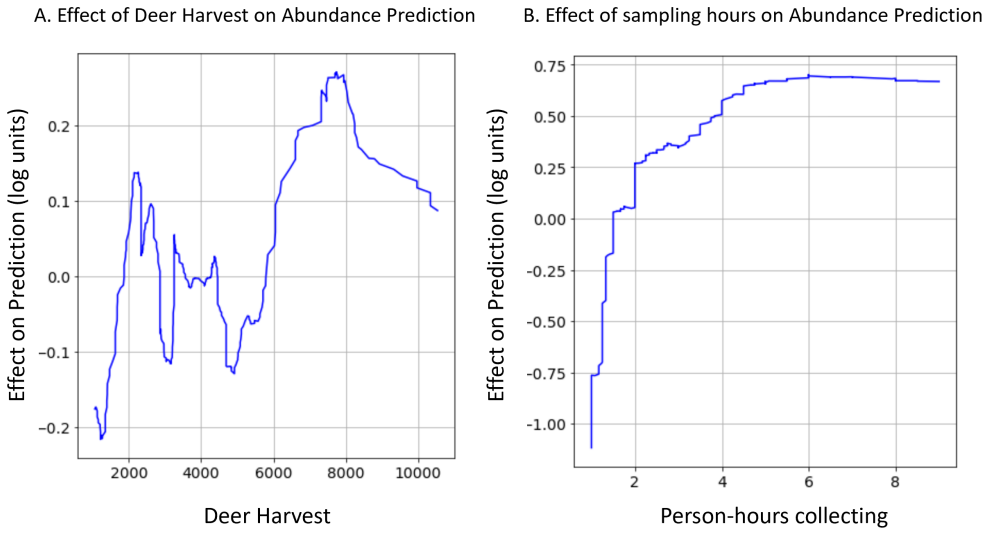

Figure 2 – Gradient boosted models identified non-linear relationships that are impractical to investigate with linear models. (A) The association between estimates of deer population size and nymphal tick abundance oscillates between having a positive effect to a negative effect. (B) The relationship between person-hours collecting hours and tick abundance is a positive but decelerating function. Data shown are the rolling average (rolling window = 50) of the impact that (A) deer density estimates or (B) tick collection effort has on tick abundance.

The impacts of non-additive interactions between environmental features on the presence of nymphal ticks were also detected in gradient boosted models. One ecologically relevant interaction demonstrates that the effect of the month in which a site is sampled on the presence of active nymphs is conditioned on the maximum temperature in June of the year before sampling (Figure 3). Although sampling month is generally highly predictive of nymphal presence due to the seasonal activity patterns of *I. scapularis* in New York State (Yuval and Spielman, 1990), ticks were more likely to be detected in the summer months (May-August) if the temperature in June of the prior year was hotter. By contrast, the probability of detecting nymphal ticks in fall months (September-December) was greater if the maximum temperature in June of the prior year was cooler. This non-additive effect was strong enough to change the month of May from being negatively associated with the presence of nymphs when June of the prior year was cooler to a positive association when this month was warmer.

The sets of environmental features used by the gradient boosted distribution and abundance models were similar to those included in linear models but were related to nymph populations

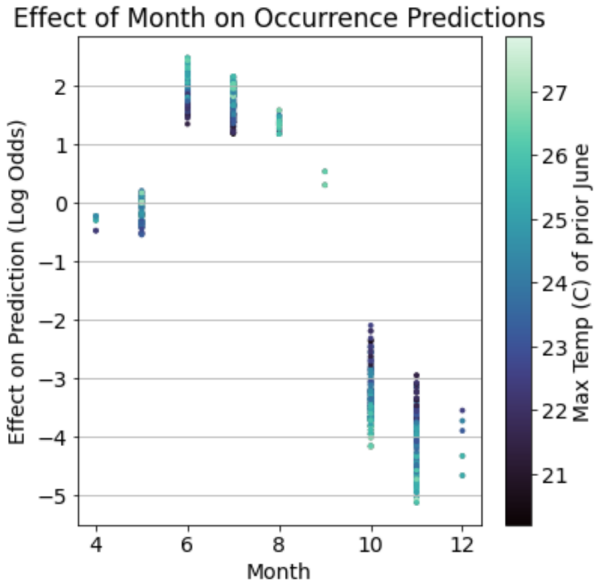

Figure 3 – Gradient boosted models detected ecologically relevant interactions between environmental features which impacts the presence of nymphal ticks. The maximum temperature in June of the year before a collection event modulates tick phenology. That is, nymphal ticks are more likely to be collected between May and August in years when the prior June was hotter while the likelihood of nymphal tick presence in September-December increases in years when the prior June was cooler.

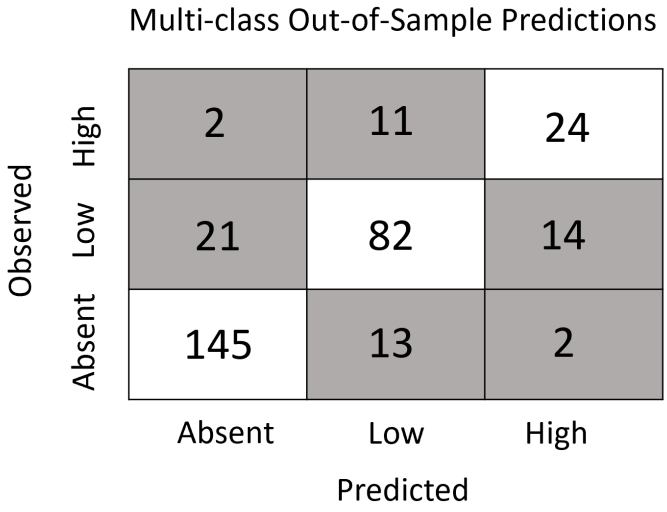

Figure 4 – The multi-class model accurately predicts both the presence and abundance of nymphs across New York State. The model accurately predicted 90.6% of sites without ticks, 70% of sites with low tick abundance (1-35), and 64.9% of sites with high tick abundance ($>$ 35). Further, most inaccurate predictions were one class apart (absent vs low or low vs high). That is, sites without nymphs were rarely predicted to have a high abundance (1.3%) and sites with high abundance were rarely predicted to have no nymphs (5.4%).

in more complex ways. Despite different feature selection processes, the two modeling frameworks frequently used identical or strongly correlated features as predictors (Supplement Table 1). However, the linear models related features to nymph populations linearly and without interaction effects, while the relationships in the gradient boosted models were always non-linear and frequently incorporated interactions. In fact, both non-linear relationships discussed above (Figure 2) involve features that were included in the previously published linear models.

The gradient boosting framework was used to produce two additional models - a multi-class and a density model - that simultaneously estimate the presence and abundance of nymphs. The multi-class model forecasts which sites will have no nymphs, low nymphal abundance (1-35), or high nymphal abundance (>35) with high accuracy, correctly classifying 80% of sites in the out-of-sample data set (Figure 4). This multi-class model predicted the presence or absence of nymphs with similar accuracy as the gradient boosted distribution model (both ≈90%) but has the additional functionality of distinguishing between two non-zero abundance classes. The novel density model predicts a continuous estimate of tick densities (ticks per collection hour) to out-of-sample data with high accuracy (R^2 = 0.42). Restricting the comparison to the subset of the out-of-sample data included in the abundance models (Figure 1B) resulted in the density model performing comparably with the linear abundance model (RMSE = 1.06 vs. 1.096; R^2 = 0.51 vs. 0.48) while retaining the added functionality of predicting the absence of nymphs. Both the multi-class and density models have similar predictive accuracy at sites that were visited for the first time in 2018 and those that had been sampled prior to 2018.

Discussion

Machine learning analyses of the recent expansion of publicly available biological and environmental data is ideal for discovering novel ecological insights and accurately forecasting the distribution and abundance of populations in nature. The gradient boosted modeling framework efficiently and accurately identifies both simple and complex ecological relationships from large data sets and produces highly accurate predictions of the demography of natural populations (Elith et al., 2008; Han et al., 2015; Ramazi et al., 2021; Wyse and Dickie, 2018). However, the theoretical advantages of gradient boosted models over traditional linear models are rarely validated using natural data sets. As a result, many ecologists rely exclusively on generalized linear models even though gradient boosted models could be more effective for exploring and interpreting data (LaRue et al., 2019; Shah et al., 2019; Sutomo et al., 2021; Walter et al., 2018). Here we demonstrate that the distribution and abundance of natural populations of *I. scapularis* ticks can be predicted with greater efficiency and accuracy with gradient boosted models than with linear models. Additionally, the gradient boosted models identified non-linear and non-additive relationships, which are difficult to detect in linear modeling frameworks, that improved predictive accuracy. These results indicate that gradient boosted models can improve both spatio-temporal forecasts and provide novel insights into the ecology of natural populations.

The gradient boosted occurrence and abundance models consistently outperformed their linear counterparts in predictive accuracy, illustrating the potential of this framework to improve predictions of ecological phenomena. When trained and tested on the same datasets as the linear models from Tran et al. (2021a), the gradient boosted models were better able to forecast the distribution and abundance of nymphs (Figure 1). Notably, the gradient boosted models outperformed their linear analogs on sites not previously sampled, suggesting that the superior predictive performance of this framework results from incorporating more precise ecological relationships rather than overfitting to previously sampled sites. However, gradient boosted models are not always expected to be the most accurate type of model for a given problem. As examples, linear models might be favored for small datasets with simpler relationships when overfitting is likely to be a problem, whereas neural networks are expected to outperform in contexts like image or speech classification (Deng et al., 2013; Hastie et al., 2001; Rawat and Wang, 2017). Nonetheless, our findings highlight gradient boosted models as a powerful but underutilized tool for predicting demographic changes in natural populations.

The gradient boosted models automatically identified complex relationships between several environmental features and the distribution and abundance of ticks. For example, these models found a non-linear relationship between deer harvest data - an estimate of deer population size - and nymphal tick abundance (Tran et al., 2021a). The non-linear relationship identified in the gradient boosted model implies that changes in deer populations are positively associated with tick abundance at some deer population sizes and negatively at others (Figure 2). This non-linear relationship may explain contradictory conclusions in previous reports in which some

identify positive relationships between deer population size and tick densities while others do not (Kugeler et al., 2016; Lewis et al., 2017; Ostfeld et al., 2006; Schulze et al., 2001; Tran et al., 2021a). Statistical models like gradient boosting do not identify the ecological mechanism underlying this relationship but do suggest avenues for further experimentation to resolve this discrepancy. Gradient boosted models also identified an interaction between climate variables that influences tick questing activity throughout summer months. Specifically, hotter temperatures in June of the year prior to tick collections alter tick phenology such that nymphal ticks are active earlier in the season (Figure 3). These results warrant further investigation into how climate change may affect seasonal activity patterns of ticks and possibly the pathogens they transmit (MacDonald et al., 2021).

Relationships between variables identified by any statistical model should be interpreted with caution. The ecological relationships included in the gradient boosted models presented here were identified using SHAP value analyses that determine the effect each variable has on model predictions (Lundberg and Lee, 2017). Thus, these relationships represent the patterns our models used to make accurate predictions but do not necessarily represent causal processes. Nevertheless, similar environmental features were detected in the gradient boosted and linear models despite using different approaches (Supplemental Table 1), adding confidence that these features are useful in forecasting tick distribution and abundance (Tran et al., 2021a). Additionally, the complex relationships involving these shared environmental features suggests that the gradient boosted framework has the potential to yield novel ecological insights, even on datasets previously analyzed with traditional statistical methods. While further experimentation is needed to clarify the biological significance of these relationships, they demonstrate the ability of the gradient boosting framework to automatically discover non-linear and interaction effects which general linear models often do not detect.

The flexibility of the gradient boosted modeling framework allowed us to build models with at least three practical advantages for both ecological interpretation and public health (De'ath, 2007). First, the multi-class and density model simultaneously predict the distribution and abundance of ticks, allowing tick population size to be estimated with a single model. Second, data pre-processing such as log-transformations is not required in the gradient boosting framework making both the predictions and error estimates more interpretable. Lastly, the density model analyzes tick density directly, a correlate of the human contact risk with a questing nymph, as opposed to the number of ticks collected which is conditioned by the sampling effort (Khatchikian et al., 2012). While it is in principle possible to achieve these advantages using generalized linear models (for an ecological example see Bah et al., 2022), the flexibility of the gradient boosting framework greatly simplified the process of implementing these multiple types of models (Natekin and Knoll, 2013).

Applying the gradient boosted modeling framework to pathogens carried by *I. scapularis* may provide additional improvements for disease risk forecasting and could identify the environmental features that correlate with human risk of contracting a *I.scapularis*-borne disease. For example, gradient boosted analyses of the distribution and abundance of ticks carrying *Borrelia burgdorferi*, *Babesia microti*, *Anaplasma phagocytophilum*, or other tick-borne pathogens are likely to identify ecological factors impacting pathogen populations and could predict the risk of encountering an infected tick. More broadly, the gradient boosted framework can improve ecological models of many infectious disease systems (Ashby et al., 2017; Fischhoff et al., 2021; Giles et al., 2018; Han et al., 2015; Solano-Villarreal et al., 2019). The rapidly expanding environmental data sets can be efficiently analyzed by gradient boosted models in order to detect ecological relationships and accurately predict disease risk in many systems, thus promoting a better understanding of natural disease systems and aiding the development of public health strategies.

Acknowledgements

Preprint version 3 of this article has been peer-reviewed and recommended by Peer Community In Ecology (https://doi.org/10.24072/pci.ecology.100532; Poisot, 2023).

Fundings

This work was supported by the NYSDOH, the National Institutes of Health (AI142572), and the Burroughs Welcome Fund (1012376).

Conflict of interest disclosure

The authors of this preprint declare that they have no financial conflict of interest with the content of this article. D Brisson is a Recommender at PCI Ecology and is on the Managing Board at PCI Evolutionary Biology.

Data, script, code, and supplementary information availability

Data used to train and validate models are from Tran et al. (2021a). Data, code for model training and evaluation, and supplement containing all features used by GBMs are available at (https://doi.org/10.17632/w8bp678m3f.2; Manley et al., 2023).

References

Abbott I, Abbott LK, Grant PR (1977). *Comparative Ecology of Galapagos Ground Finches (*Geospiza Gould*): Evaluation of the Importance of Floristic Diversity and Interspecific Competition*. *Ecological Monographs* **47**, 151–184. https://doi.org/10.2307/1942615.

Ashby J, Moreno-Madriñán MJ, Yiannoutsos CT, Stanforth A (2017). *Niche Modeling of Dengue Fever Using Remotely Sensed Environmental Factors and Boosted Regression Trees*. *Remote Sensing* **9**. https://doi.org/10.3390/rs9040328.

Austin MP, Cunningham RB, Fleming PM (1984). *New approaches to direct gradient analysis using environmental scalars and statistical curve-fitting procedures*. *Vegetation* **55**, 11–27. https://doi.org/10.1007/BF00039976.

Austin MP, Nicholls AO, Margules CR (1990). *Measurement of the Realized Qualitative Niche: Environmental Niches of Five Eucalyptus Species*. *Ecological Monographs* **60**, 161–177. https://doi.org/10.2307/1943043.

Bah MT, Grosbois V, Stachurski F, Muñoz F, Duhayon M, Rakotoarivony I, Appelgren A, Calloix C, Noguera L, Mouillaud T, Andary C, Lancelot R, Huber K, Garros C, Leblond A, Vial L (2022). *The Crimean-Congo haemorrhagic fever tick vector* Hyalomma marginatum *in the south of France: Modelling its distribution and determination of factors influencing its establishment in a newly invaded area*. *Transboundary and Emerging Diseases* **69**. https://doi.org/10.1111/tbed.14578.

Becker EA, Carretta JV, Forney KA, Barlow J, Brodie S, Hoopes R, Jacox MG, Maxwell SM, Redfern JV, Sisson NB, Welch H, Hazen EL (2020). *Performance evaluation of cetacean species distribution models developed using generalized additive models and boosted regression trees*. *Ecology and Evolution* **10**, 5759–5784. https://doi.org/10.1002/ece3.6316.

Bentéjac C, Csörgő A, Martínez-Muñoz G (2021). *A comparative analysis of gradient boosting algorithms*. *Artificial Intelligence Review* **54**, 1937–1967. https://doi.org/10.1007/s10462-020-09896-5.

Burgdorfer W, Barbour AG, Hayes SF, Benach JL, Grunwaldt E, Davis JP (1982). *Lyme Disease - a Tick-Borne Spirochetosis? Science* **216**, 1317–1319. https://doi.org/10.1126/science.7043737.

Cawley GC, Talbot NLC (2010). *On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation*. *Journal of Machine Learning Research* **11**, 2079–2107.

Cutler DR, Edwards Jr. TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007). *Random Forests for Classification in Ecology*. *Ecology* **88**, 2783–2792. https://doi.org/https://doi.org/10.1890/07-0539.1.

De'ath G (2007). *Boosted Trees for Ecological Modeling and Prediction*. *Ecology* **88**, 243–251. https://doi.org/10.1890/0012-9658(2007)88[243:btfema]2.0.co;2.

De'ath G, Fabricius KE (2000). *Classification and regression trees: a powerful yet simple technique for ecological data analysis*. *Ecology* **81**, 3178–3192. https://doi.org/10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2.

Deng L, Hinton G, Kingsbury B (2013). *New types of deep neural network learning for speech recognition and related applications: an overview*. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 8599–8603. https://doi.org/10.1109/ICASSP.2013.6639344.

Elith J, Leathwick JR, Hastie T (2008). *A working guide to boosted regression trees*. *Journal of Animal Ecology* **77**, 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x.

Elith J, H. Graham C, P. Anderson R, Dudík M, Ferrier S, Guisan A, J. Hijmans R, Huettmann F, R. Leathwick J, Lehmann A, Li J, G. Lohmann L, A. Loiselle B, Manion G, Moritz C, Nakamura M, Nakazawa Y, McC. M. Overton J, Townsend Peterson A, J. Phillips S, et al. (2006). *Novel methods improve prediction of species' distributions from occurrence data*. *Ecography* **29**, 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x.

Elith J, Leathwick JR (2009). *Species Distribution Models: Ecological Explanation and Prediction Across Space and Time*. *Annual Review of Ecology, Evolution, and Systematics* **40**, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159.

Escobar LE, Qiao H, Cabello J, Peterson AT (2018). *Ecological niche modeling re-examined: A case study with the Darwin's fox*. *Ecology and Evolution* **8**, 4757–4770. https://doi.org/10.1002/ece3.4014.

Farley SS, Dawson A, Goring SJ, Williams JW (2018). *Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions*. *BioScience* **68**, 563–576. https://doi.org/10.1093/biosci/biy068.

Fischhoff IR, Castellanos AA, Rodrigues JPGLM, Varsani A, Han BA (2021). *Predicting the zoonotic capacity of mammals to transmit SARS-CoV-2*. *Proceedings of the Royal Society B: Biological Sciences* **288**, 20211651. https://doi.org/10.1098/rspb.2021.1651.

Friedman JH (2001). *Greedy function approximation: A gradient boosting machine. The Annals of Statistics* **29**, 1189–1232. https://doi.org/10.1214/aos/1013203451.

Giles JR, Eby P, Parry H, Peel AJ, Plowright RK, Westcott DA, McCallum H (2018). *Environmental drivers of spatiotemporal foraging intensity in fruit bats and implications for Hendra virus ecology*. *Scientific Reports* **8**, 9555. https://doi.org/10.1038/s41598-018-27859-3.

Grinsztajn L, Oyallon E, Varoquaux G (2022). *Why do tree-based models still outperform deep learning on typical tabular data?* **35**. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, 507–520.

Han BA, Schmidt JP, Bowden SE, Drake JM (2015). *Rodent reservoirs of future zoonotic diseases*. *Proceedings of the National Academy of Sciences* **112**, 7039–7044. https://doi.org/10.1073/pnas.1501598112.

Harvey PH, Clutton-Brock TH, Mace GM (1980). *Brain size and ecology in small mammals and primates. Proceedings of the National Academy of Sciences of the United States of America* **77**, 4387–4389.

Hastie T, Tibshirani R, Friedman J (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second Edition. New York: Springer.

Khatchikian CE, Prusinski M, Stone M, Backenson PB, Wang IN, Levy MZ, Brisson D (2012). *Geographical and environmental factors driving the increase in the Lyme disease vector* Ixodes scapularis. *Ecosphere* **3**, 85. https://doi.org/10.1890/ES12-00134.1.

Kleiber M (1947). *Body size and metabolic rate*. *Physiological Reviews* **27**, 511–541. https://doi.org/10.1152/physrev.1947.27.4.511.

Kugeler KJ, Jordan RA, Schulze TL, Griffith KS, Mead PS (2016). *Will Culling White-Tailed Deer Prevent Lyme Disease? Zoonoses and Public Health* **63**, 337–345. https://doi.org/10.1111/zph.12245.

LaRue M, Salas L, Nur N, Ainley D, Stammerjohn S, Barrington L, Stamatiou K, Pennycook J, Dozier M, Saints J, Nakamura H (2019). *Physical and ecological factors explain the distribution of Ross Sea Weddell seals during the breeding season*. *Marine Ecology Progress Series* **612**, 193–208. https://doi.org/10.3354/meps12877.

Levin SA (1998). *Ecosystems and the Biosphere as Complex Adaptive Systems*. *Ecosystems* **1**, 431–436. https://doi.org/10.1007/s100219900037.

Lewis JS, Farnsworth ML, Burdett CL, Theobald DM, Gray M, Miller RS (2017). *Biotic and abiotic factors predicting the global distribution and population density of an invasive large mammal*. *Scientific Reports* **7**, 44152. https://doi.org/10.1038/srep44152.

Lundberg SM, Lee SI (2017). *A Unified Approach to Interpreting Model Predictions*. In: *Procedings of the 31st international conference on neural information processing systems*. Vol. 30. Curran Associates, Inc., pp. 4768–4777. https://doi.org/10.48550/arXiv.1705.07874.

MacDonald H, Akçay E, Brisson D (2021). *The role of host phenology for parasite transmission*. *Theoretical Ecology* **14**, 123–143. https://doi.org/10.1007/s12080-020-00484-5.

Manley W, Tran T, Prusinski M, D B (2023). *Modeling Tick Populations: An Ecological Test Case for Gradient Boosting Trees*. *Mendeley Data*, *2*. https://doi.org/10.17632/w8bp678m3f.2.

McCullagh P, Nelder J (1989). *Generalized Linear Models*. Second Edition. CRC Press.

McLain DK, Moulton MP, Redfearn TP (1995). *Sexual Selection and the Risk of Extinction of Introduced Birds on Oceanic Islands*. *Oikos* **74**, 27–34. https://doi.org/10.2307/3545671.

Naghibi SA, Pourghasemi HR (2015). *A Comparative Assessment Between Three Machine Learning Models and Their Performance Comparison by Bivariate and Multivariate Statistical Methods in Groundwater Potential Mapping*. *Water Resources Management* **29**, 5217–5236. https://doi.org/10.1007/s11269-015-1114-8.

Natekin A, Knoll A (2013). *Gradient boosting machines, a tutorial*. *Frontiers in Neurorobotics* **7**. https://doi.org/10.3389/fnbot.2013.00021.

Olden J, Lawler J, Poff N (2008). *Machine Learning Methods Without Tears: A Primer for Ecologists*. *The Quarterly Review of Biology* **83**, 171–193. https://doi.org/10.1086/587826.

Ostfeld RS, Canham CD, Oggenfuss K, Winchcombe RJ, Keesing F (2006). *Climate, Deer, Rodents, and Acorns as Determinants of Variation in Lyme-Disease Risk*. *PLoS Biology* **4**, e145. https://doi.org/10.1371/journal.pbio.0040145.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Müller A, Nothman J, Louppe G, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011). *Scikit-learn: Machine Learning in Python*. *Journal of Machine Learning Research* **12**, 2825–2830.

Poisot T (2023). *Gradient Boosted Trees can deliver more than accurate ecological predictions*. *Peer Community in Ecology*, 100532. https://doi.org/10.24072/pci.ecology.100532.

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*. *Journal of Medical Entomology* **51**, 226–236. https://doi.org/10.1603/ME13101.

Qiao H, Soberón J, Peterson AT (2015). *No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation*. *Methods in Ecology and Evolution* **6**, 1126–1136.

Ramazi P, Kunegel-Lion M, Greiner R, Lewis MA (2021). *Predicting insect outbreaks using machine learning: A mountain pine beetle case study*. *Ecology and Evolution* **11**, 13014–13028. https://doi.org/10.1002/ece3.7921.

Rammer W, Seidl R (2019). *Harnessing Deep Learning in Ecology: An Example Predicting Bark Beetle Outbreaks*. *Frontiers in Plant Science* **10**. https://doi.org/10.3389/fpls.2019.01327.

Raschka S (2018). *MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack*. *Journal of Open Source Software* **3**. https://doi.org/10.21105/joss.00638.

Rawat W, Wang Z (2017). *Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review*. *Neural Computation* **29**, 2352–2449. https://doi.org/10.1162/NECO_a_00990.

Root T (1988). *Energy Constraints on Avian Distributions and Abundances*. *Ecology* **69**, 330–339. https://doi.org/10.2307/1940431.

Rudin C (2019). *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*. *Nature Machine Intelligence* **1**, 206–215. https://doi.org/10.1038/s42256-019-0048-x.

Ryo M, Angelov B, Mammola S, Kass JM, Benito BM, Hartig F (2021). *Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models*. *Ecography* **44**, 199–205. https://doi.org/10.1111/ecog.05360.

Schulze TL, Jordan RA, Hung RW (2001). *Potential Effects of Animal Activity on the Spatial Distribution of* Ixodes scapularis *and* Amblyomma americanum *(Acari: Ixodidae)*. *Environmental Entomology* **30**, 568–577. https://doi.org/10.1603/0046-225X-30.3.568.

Shabani F, Kumar L, Ahmadi M (2016). *A comparison of absolute performance of different correlative and mechanistic species distribution models in an independent area*. *Ecology and Evolution* **6**, 5973–5986. https://doi.org/10.1002/ece3.2332.

Shah MM, Krystosik AR, Ndenga BA, Mutuku FM, Caldwell JM, Otuka V, Chebii PK, Maina PW, Jembe Z, Ronga C, Bisanzio D, Anyamba A, Damoah R, Ripp K, Jagannathan P, Mordecai EA, LaBeaud AD (2019). *Malaria smear positivity among Kenyan children peaks at intermediate temperatures as predicted by ecological models*. *Parasites & Vectors* **12**, 288. https://doi.org/10.1186/s13071-019-3547-z.

Solano-Villarreal E, Valdivia W, Pearcy M, Linard C, Pasapera-Gonzales J, Moreno-Gutierrez D, Lejeune P, Llanos-Cuentas A, Speybroeck N, Hayette MP, Rosas-Aguirre A (2019). *Malaria risk assessment and mapping using satellite imagery and boosted regression trees in the Peruvian Amazon*. *Scientific Reports* **9**, 15173. https://doi.org/10.1038/s41598-019-51564-4.

Spielman A, Clifford CM, Piesman J, Corwin MD (1979). *Human Babesiosis on Nantucket Island, USA: Description of the Vector,* Ixodes Dammini, *N. Sp. (Acarina: Ixodidae)*. *Journal of Medical Entomology* **15**, 218–234. https://doi.org/10.1093/jmedent/15.3.218.

Stephens PR, Pappalardo P, Huang S, Byers JE, Farrell MJ, Gehman A, Ghai RR, Haas SE, Han B, Park AW, Schmidt JP, Altizer S, Ezenwa VO, Nunn CL (2017). *Global Mammal Parasite Database version 2.0*. *Ecology* **98**, 1476. https://doi.org/10.1002/ecy.1799.

Sutomo, Yulia E, Iryadi R (2021). *Kirinyuh (*Chromolaena odorata*): species distribution modeling and the potential use of fungal pathogens for its eradication*. *IOP Conference Series: Earth and Environmental Science* **762**, 012023. https://doi.org/10.1088/1755-1315/762/1/012023.

Telford SR, Dawson JE, Katavolos P, Warner CK, Kolbert CP, Persing DH (1996). *Perpetuation of the agent of human granulocytic ehrlichiosis in a deer tick-rodent cycle*. *Proceedings of the National Academy of Sciences* **93**, 6209–6214. https://doi.org/10.1073/pnas.93.12.6209.

Tilman D, Wedin D, Knops J (1996). *Productivity and sustainability influenced by biodiversity in grassland ecosystems*. *Nature* **379**, 718–720. https://doi.org/10.1038/379718a0.

Tran T, Porter WT, Salkeld DJ, Prusinski MA, Jensen ST, Brisson D (2021b). *Estimating disease vector population size from citizen science data*. *Journal of The Royal Society Interface* **18**, 20210610. https://doi.org/10.1098/rsif.2021.0610.

Tran T, Prusinski MA, White JL, Falco RC, Vinci V, Gall WK, Tober K, Oliver J, Sporn LA, Meehan L, Banker E, Backenson PB, Jensen ST, Brisson D (2021a). *Spatio-temporal variation in environmental features predicts the distribution and abundance of* Ixodes scapularis. *International Journal for Parasitology* **51**, 311–320. https://doi.org/10.1016/j.ijpara.2020.10.002.

Walter T, Zink R, Laaha G, Zaller JG, Heigl F (2018). *Fox sightings in a city are related to certain land use classes and sociodemographics: results from a citizen science project*. *BMC Ecology* **18**, 50. https://doi.org/10.1186/s12898-018-0207-7.

Wyse SV, Dickie JB (2018). *Taxonomic affinity, habitat and seed mass strongly predict seed desiccation response: a boosted regression trees analysis based on 17539 species*. *Annals of Botany* **121**, 71–83. https://doi.org/10.1093/aob/mcx128.

Yee TW, Mitchell ND (1991). *Generalized additive models in plant ecology*. *Journal of Vegetation Science* **2**, 587–602. https://doi.org/10.2307/3236170.

Yuval B, Spielman A (1990). *Duration and Regulation of the Developmental Cycle of* Ixodes dammini *(Acari: Ixodidae). Journal of Medical Entomology* **27**, 196–201. https://doi.org/10.1093/jmedent/27.2.196.

Supplementary Data

Supplemental Table 1: Most Predictive Ecological Features from Gradient Boosted Occurrence and Abundance Models compared to Linear Counterparts

Model	GBM Occurrence	GLM Occurrence	GBM Abundance	GLM Abundance
Physical Habitat	Longitude (+, NL), Distance to nearest road (-, NL)	Latitude (+), Elevation (-), Distance to nearest road (+), Road type of nearest road (NL), Indicator of critical zone (-)	Latitude (-, NL), Longitude (+, NL)	Latitude (-), Longitude (+), Elevation (NL), Forest (-), Distance to nearest hydrography feature (-)
Vapor Pressure	Maximum Jan 2 years prior (-, NL), Minimum Oct 2 years prior (NL), Maximum Oct 1 year prior (+, NL), Maximum Jan (-, NL), Minimum June (-, NL), Minimum October (+, NL)	Minimum Jan 1 year prior (-)		Maximum October 2 years prior (+), Minimum October 2 years prior (-, NL)
Temperature	Mean differential Jan 2 years prior (+, NL, IE), Degree days above 0 C spring-summer 1 year prior (+, NL, IE), Degree days above 0 C spring 1 year prior (+, NL), Maximum June 1 year prior (+, NL, IE)	Degree days above 0 C spring 2 years prior (-), Degree Days below 0 C winter 1 year prior (+), Degree days above 0 C spring-summer 1 year prior (+)		Degree days above 0 C spring-summer 1 year prior (+)
Day of Collection	Person-hours collecting (+, NL), Month (NL)	Person-hours collecting (+), Month (NL), Local Temperature (+), Wet (-)	Person-hours collecting (+, NL), Week (NL)	Person-hours collecting (+), Month (NL)
Miscellaneous		Deer harvest (-)	Deer harvest (NL)	Deer harvest (+)

Top 15 most predictive features from the gradient boosted occurrence model and all features from the other models are included.
(-) = negative relationship, (+) = positive relationship, NL = nonlinear relationship, IE = interaction effect

Supplemental Table 2: Summary of Model Characteristics

Model	Sites Predicted	Target Variable	Accuracy Metrics for Out of Sample Test	GLM Analog
GBM Distribution	All sites	Binary (Nymphs Present or Absent)	Accuracy, Sensitivity, Specificity	GLM Distribution Model
GBM Abundance	Sites with Nymphs	Log-transformed Nymph Abundance	RMSE, R^2, Categorical Accuracy	GLM Abundance Model
GBM Multi-Class	All sites	Three Abundance Classes of Nymphs	Accuracy	N/A
GBM Density	All sites	Nymph Abundance/ Sampling Hour	RMSE, R^2	N/A

Model characteristics of all four gradient boosted models are included.