Response to Recommender

Additional clarifications in the methodology are required

Both reviewers are positive about the study, and their comments generally deal with (i) a better justification of methodological choices, and (ii) more transparency in the interpretation of the results, in particular when it comes to discussing causal relationships inferred with the GBM.

-*We thank you for facilitating the review process. Responding to the comments resulted in a much stronger manuscript. (i) We have included additional justification for using gradient boosting rather than other machine learning methods and for why the results are unlikely to be contingent on our specific methodological choices during feature selection. (ii) We have also added a more nuanced discussion concerning the inference of ecological relationships using GBMs. These methods identify important hypotheses about correlations in the data but we recognize that, without experimentation, we cannot infer their ecological meaning. We now better emphasize that these relationships are examples supporting the point that GBMs are a powerful tool for automatically detecting complex relationships from large ecological datasets.*

I would encourage the authors to pay attention to the comment from reviewer 2 about the need to repliacte the original formulation of models coming from other publications. This is an important point in order to set up a fair comparison of models -- this is not to say that deviations are inadmissible, but they must be strongly justified.

-*We apologize for the lack of clarity in our original manuscript as our models do replicate the original formulation of the cited linear models. We have clarified our methodology throughout our text and have included an additional supplemental table to add further clarity.*

As some of the comments from the reviewers will lead to new (or updated) analyses, I am anticipating that I will send the revised version to review; with this in mind, please write a detailed response to the reviewers, in order to facilitate the next round of comments.

-*Below we provide a point-by-point response to reviewers. Many thanks for your help with this process.*

Response to Reviewer 1

This paper deals with the application of machine learning modeling to ecological data and its comparison with more classical linear modeling. The topic is relevant and the article is generally clear with interesting results. I have however some points that should be addressed or amended to my opinion.

Major points:

- The authors do not justify the use of gradient boosted trees method. There are other techniques that could be applied to such a dataset and the reader should be shortly informed about the advantages and drawbacks of gradient boosted trees with respect to other major methods.

We agree with this point and have added a brief discussion the advantages and disadvantages of gradient boosted trees in the context of some other major methods. We now better justify our use of gradient boosted trees (p2: L4-9). Additionally, we now discuss some drawbacks compared to other ML methods in the discussion on P8L32-37.

- It is not clear why the linear model could not take into account tick density data, which seems to constitute an asset of machine learning model. Even though this is not the subject of the paper, there are examples of zero inflated generalized linear models (with Poisson distribution : see Bah et al., 2022, DOI: 10.1111/tbed.14578, but negative binomial could also be considered) which consider both occurrence and abundance. Data with no occurrence could also be used by applying a log(y+1) transformation. Therefore, these kinds of linear model may lower the advantage of using machine learning techniques and this should be acknowledged

We apologize that the language in our prior draft implied a combined distribution/abundance model could not be done in the linear framework. We agree this is possible to do in the framework with methods such as the ones you have cited. We made multiple changes to clarify this point.

Intro: P2L30-34: Rather than implying that linear models could not have possibly simultaneously modeled distribution and abundance, we make it clear that this limitation was due to the specific assumptions made about the data distributions in Tran et al. 2021.

Methods: P3L14-16: When we describe why distribution and abundance models were built separately in the linear framework, we now attribute this to the specific assumptions made by the cited linear models rather than an inherent limitation to the framework.

Discussion: P9 L27-30: We are now more careful to not imply that general linear models could not have simultaneously modeled distribution and abundance. We also explicitly make the point that it is in principle possible to achieve these model types using GLMs but the flexibility of GBMs made their implementation more practical (no need to fit specific data distributions for different model types).

- My main concern lies in the use of machine learning methods to help interpreting ecological interactions. It appears straightforward that machine learning significantly improves the predictive capacity of models and this is shown by the present paper. I am not convinced by the use of GBM to investigate the influence of environmental features with the present study.

An increase in tick density with deer density, with a kind of linear-plateau relationship, seems relevant. But the variations of this influence for intermediate deer harvest does not appear to be based on biological grounds. The same applies to the influence of the temperature in June of the year before which have not been accounted for in other ecological studies on ticks, to my knowledge. These variables also arise from the linear model, which gathers many variables to my knowledge. In any cases I do not see the advantage of using GBM in this context. The authors acknowledge that GBM could point out particular issues to be addressed in detail, in the discussion (as linear models could also do) but they should be more cautious throughout the text.

We agree with the reviewer and apologize for lack of clarity. We agree that we cannot conclude that the non-linear correlations we identify are biologically relevant. We now better explain that these correlations can be hypotheses that can be experimentally tested (P8L38-P9L19). Further, we clarify that the advantage of gradient boosted models is the ability to explore non-linear and interaction effects automatically (without a priori *expectations as needed in classical linear models). Exploring all possible non-linear and interaction effects is impractical in the GLM framework due to the size of the dataset. Instances where the GLMs and GBMs identified the same features but different relationship shapes demonstrate the potential of GBMs to identify complex ecological relationships that have been missed from prior analyses with GLMs. Throughout the manuscript, we are now more careful to point out that we do not know the biological significance of the relationships discussed, but that they serve as demonstrations for GBMs ability to automatically incorporate complex effects, which are often missed by GLMs.*

- How does the quality of fitting vary with the maximal number of environmental features involved ? Why have you fixed this number at a value of 30 ? If you lower this number, does it have an influence on the fitting ?

This number was fixed at a maximum of 30 features for three main reasons. First, the number of model fits required in the feature selection algorithm grows exponentially with the maximal value resulting in a computational burden. Second, choosing too many features can lead to overfitting. Third, we wanted to keep the number of features in the same ballpark as used for the General Linearized Models. All of our models included fewer than 25 features suggesting that raising the maximal value would not change results. We do expect that a dramatically lower number would impact predictive power due to information loss.

Other points:

- P6 L21-22: the difference between RMSE and R2 for both models is quite low. I do not think that we could say that the density model is outperforming the abundance linear model.

We agree saying the density model outperformed the linear model here was a bit too strong for the performance difference, we now claim that they have comparable performance. See (P8 L4-7).

- In the discussion, I don't think it is worth getting into the detail of model results (P8 L21-27).

We agree, this was repetitive and have replaced it with a more detailed discussion of the predictive capacities of GBMs compared to other models. See (P8 L29-37).

Response to Reviewer 2

Manley et al present an interesting comparison of linear models and boosted regression trees in modelling tick distributions. The paper is succinct and nicely carried out, and I think it will make a good contribution to the literature. It builds largely on earlier published findings from the same group, which linked environmental variables with tick distributions. This paper therefore represents an improvement on their earlier findings as well as a case study investigating the utility of machine learning approaches compared to linear models. I'm not an expert in the machine learning approaches involved but the models appear to have been designed and fitted well, and the results make sense to me.

Thank you, we hope it is a useful contribution to the field.

My one comment is that the authors don't fully justify a few aspects of their modelling approach, which makes it unclear how direct the linear:BRT comparison is. If the authors could add a bit more detail and justification about the comparisons between the two (and the compromises involved in making the comparisons) that would help to critically assess the results.

· Why did the abundance approach use a multiclass model rather than actually predicting count? Would the results be the same?

We apologize that this was not clear in the prior submission. The linear models of tick abundances were trained and predicted tick counts and these outputs were then converted into classes. We used the same approach with the gradient boosted models of tick abundances so as to directly compare the models using the accuracy metrics published in the linear model paper (Tran et al 2021). We have clarified this in the resubmitted manuscript at p3(L41-45), p4(L37-41), and in supplemental table 2.

· Did the linear models from the earlier paper use the same multiclass approach? If not, does that not introduce nonequivalences into the comparison of the models? I was confused about exactly what the response variables and model constructions were when reading both papers side by side; it seems like the previous approach used a linear model that was then categorized into the abundance classes after the predictions? If so, why not do the same here? The paper could do with a table or similar in the supplement that very simply presents this information side by side, detailing what the response variables are and which models they were incorporated in.

We apologize this was not clear in the prior submission, we used the same response variables and model constructions as the earlier paper. Please see the above comment for how this was clarified in the current submission.

· If the linear models in the previous paper used different response variables/construction then they should ideally be repeated in this paper with identical formulations if possible. Otherwise it's not totally clear whether the values reported are valid comparisons, or whether apples are being compared to oranges in some ways. I have no doubt that the BRTs are better at predicting the observed trends compared to linear models regardless, but it could influence the difference in predictive ability and therefore the strength of the argument. I would explicitly number and name the models (Model 1A, 1B, 2A, 2B etc) to make them easy to refer to in each area of the paper.

We hope with the clarifications in our above responses it is clear that our models had the same formulations as the linear models and therefore are valid to compare. We agree model names are

useful for clarification and now refer to them as either distribution, abundance, multi-class, or density model. These names are presented with the model attributes in supplemental table 2.

· If the linear models are repeated the authors could use a mixture model like a zero-inflated or hurdle model to examine presence and abundance in the same model rather than separating them (and log-transforming them wouldn't be necessary with a negative binomial model), but for the purposes of this study it might be simplest just to log-transform the response when running the BRTs and running a more comparable model.

We apologize this was not clear in the prior submission, we used the same response variables and model constructions as the earlier paper. Please see the above responses for how this was clarified in the current submission.

Minor points:

· Figure 1: map lines appear to be on top of points; the points would look better as the top layer with the map underneath.

We agree this looks better, the points are now the top layer with the map underneath.

· Figure 2B: This is possibly not the best choice of a relationship to display here, as this relationship could be quite easily approximated by a linear model applied to transformed data. If the authors are determined to stick with this relationship, it might be worth mentioning this.

Our intention here was not to assert that the relationship from this figure could not in principle be identified by linear models, which we agree it could have been. Rather, we are using this relationship to demonstrate the advantage GBMs have in automatically incorporating nonlinearities, which leads them to find nonlinearities that are often missed by GLMs. We believe the presented relationship makes this point well because the GLMs modeled this same predictor variable but missed its nonlinear effect on tick abundance. We have now clarified the point this example is used to make at P9(L17-19).

· The line "deer harvest data, an estimate of deer population size, and nymphal tick abundance (Tran et al., 2021a)." At first reading it's unclear if this is three variables or whether the middle one a description of the first. I'd suggest delineating the middle clause with brackets or an em-dash rather than commas.

We agree this was unclear, we now have delineated the middle clause with dashes.