Reviewer #1: Revision of "Performance evaluation and hyperparameter tuning of statistical and machine-learning models using spatial data"  
  
In this study, the authors analyze the effect of spatial cross-validation strategies and the effect of spatial hyperparameter tuning for several parametric and non-parametric algorithms. The comparisons are performed at the example of a prediction task aiming at the classification of a forest disease.  
  
The paper is highly relevant for the predictive modelling community due to the widely underestimated problem of overfitting in spatial prediction tasks. I read it with highest interest and recommend it for publication after the issues and questions listed below are clarified. In summary, my major concerns are that the objectives and hypotheses of this study are not made very clear which remains a problem throughout the manuscript. Also the key challenges in spatial ML applications that have motivated this study are not well elaborated in the text. Finally some of the conclusions are not supported by the results of this study.   
  
I hope the authors of the manuscript find the revisions helpful. I'm looking forward to read the revised version of this manuscript.  
  
-----------------------------------  
  
MAJOR COMMENTS  
  
1) Clearly state your objectives and hypotheses. It remains unclear what exactly should be achieved or explored with this study. For example it is unclear why you use different ML algorithms. I assume your objective is to compare the effect of hyperparameter tuning for several algorithms. What is your hypothesis in this case? I assume it is that different algorithms have a different sensitivity to spatial hyperparameter tuning and to spatial vs non spatial CV in general? I suggest you use bullet points to explicitly highlight each of your aims (i.e. sensitivity to hyperparameter tuning, spatial vs non spatial cross validation, comparison of the sensitivity to tuning between algorithms, nonparametric vs parametric) and the corresponding hypothesis.  
  
2) The description of the case study needs a bit more detail and a short but clear summary of  Iturritxa et al. I suggest including a first subsection in chapter 2, called something like "Summary of the prediction task" where you clearly explain the aim and setting of this project. It further remains unclear how you assume that your newly chosen predictors matter (why is hail damage or soil type important to map tree infection?). From a technical perspective: How does spatial autocorrelation becomes present in this example?   
  
3) Clearly explain the challenges of ML based spatial mapping in the introduction. Especially the problem of spatial autocorrelation. This is the crucial problem leading to your idea to test the effect of hyperparameter tuning and hence it deserves a proper description. Also make clear what the problem with a "standard cross validation" is in comparison to spatial CV (e.g. why does it lead to overoptimistic results?). Why does this problem leads to the need of spatial hyperparameter tuning? Without making these points clear it is hard to understand the relevance of your findings.  
  
4) The structure of the manuscript is confusing in some parts. I suggest leaving the long description of the algorithms out (page 14-16) as this can be easily referred to. This would help reducing the content to the essential parts. page 14 paragraph 3.4 needs to be partly shifted forwards in the manuscript. You already mentioned quite a few R packages earlier on, so it makes sense to mention right at the start of your methods chapter what software environment you're using. The discussion includes a lot of side-discussions (e.g. model interpretability, variable selection,moothness of the predictions). I strongly recommend to stick to the plot of your study rather than go into discussions about what else might be relevant.  
  
5) The conclusions do not always fit to the data. It is unclear why you recommend to use spatial partitioning for hyperparameter tuning of spatial data when the results indicate no major effect of spatial vs non-spatial hyperparameter tuning. In the same context you write that e.g. mtry needs to be tuned in a spatial way even though the performance remains the same. You justify this by the finding that the mtry values are not identical between spatial and non spatial tuning. However, the performance is the same so what effect does it have? To approach this you could compare the differences in the maps of predicted probabilities of the model using the spatially tuned hyperparameters with the non spatially tuned model. Are there patterns that indicate that the way of tuning makes a difference? Are there even differences that are beyond randomness? If not I think it is more appropriate to conclude that contrary to your hypothesis spatial hyperparameter tuning didn't make a  
difference in your case study.  
On the other hand, some important results are not highlighted in the conclusions, e.g. that the risk of overoptimistic model results in the case of non spatial CV is present for all algorithms. Please clearly align your aims/hypotheses with your conclusions (see also the first major comment).  
  
6) I think figure 6 points on an interesting finding that you did not mention: apparently the choice of the algorithm is more important than spatial vs non spatial hyperparameter tuning, while regardless of the algorithm, a spatial cross validation is crucial. Consider including this finding.  
  
7) The usage of mtry = 1 is interesting. It means that RF has no chance to skip uninformative predictors. All predictors are forced to be used even if they are just random noise. Is this desirable?   
  
8) Is it a fair comparison if you set arbitrary values as a default for the "no tuning" models? Can you justify the choice of your default values?  
  
9) It is unclear why you include GAM and GLM in your study when your main focus is on hyperparameter tuning. I see two options here. Either make explicitly clear why you include them in your study and also explicitly discuss what you learned from the outcome of these two models. Or drop them completely from your manuscript which I think might be appropriate as they don't have hyperparameters to tune which however in the current version seems to be the focus of the manuscript.  
  
-----------------------------------  
  
MINOR COMMENTS  
  
1) throughout the manuscript: make sure that it's always clear that the results of the non-spatial CV are not a valid indicator for model performance because it does not validate the performance to predict beyond the neighborhood of training locations.  
  
2) The title of the manuscript is not very clear. Both, machine learning algorithms as well as parametric models can be referred to as "statistical". I recommend to either call it "parametric and machine-learning models" or just "spatial predictive models". Further, it might make sense to change the wording to "performance evaluation and sensitivity to hyperparameter tuning" rather than just naming the key words.  
  
3) page 1, line 45-48: I think these are two things that need to be separated: We need cross-validation methods to evaluate the model performance and to obtain non-biased performance estimates. Hyperparameter tuning is important to receive robust or biased-reduced models but we're not performing tuning studies to obtain performance estimates. This should be separated by language.  
  
4) page 2, line 3-5: As you submit to ecological modelling it is rather of interest where statistical learning is used in the context of ecology.   
  
5) page 2, first paragraph: I recommend shifting the plot: First start with the motivation from a ecological perspective ("Spatial predictions are of great importance...") which leads to your introduction to the case study (page 3 lines 21-31) as a typical spatial mapping example. The case study is currently poorly linked. Continue with the methodological approaches to face the challenges that exist with such modelling tasks (spatial autocorrelation, complex relationships etc).  
  
6) page 2 in the abstract: I suggest changing this paragraph to something like this: "Results show that GAM and RF (AUROC estimates 0.705) outperform all other methods in predictive accuracy. The AUROC differences between the bias-reduced (spatial cross-validation) and overoptimistic (non-spatial cross-validation) performance estimates of the GAM and RF are 0.167 (24 %) and 0.213 (30 %), respectively. Though the sensitivity to hyperparameter tuning differed between the ML algorithms, there were in most cases no significant differences between spatial and non-spatial partitioning for hyperparameter tuning."  
  
7) page 4, line 56: define "bias reduced". When is a model biased in this context?  
  
8) page 4, line 65: Here is an explanation missing why not only model performance but also the hyperparameters might need to be assessed using spatial resampling.  
  
9) page 4, line 72: Include a line break  
  
10) Page 5 line 88: You mention the problem of uneven distribution in the dataset. How do you deal with this in your analysis?  
  
11) Page 6, lines 92-95: Are the climatic variables long-term averages?  
  
12) Page 7, lines 122-124: Shift this to the discussion   
  
13) Page 8, line 167: How are spatially disjoint subsets determined?  
  
14) page 17 line 355 and Figure 4: You say that SVM shows the highest tuning effects. But in Fig. 4 it looks like the effect is even higher for BRT.   
I also wonder how WKNN can be better without tuning? Is there a lot of randomness in the models? In general, how is it possible that the performance sometimes decreases with tuning iterations (e.g. BRT at 300 iterations or SVM at 50 iterations). Maybe providing standard error bars might help in assessing the relevance of small differences.  
  
15) Page 20, line 383-388: I agree that hyperparameter tuning is expensive in computation time. You use a comparably small dataset but if large datasets are used it is nearly impossible to perform a such extensive tuning study. I think it is worth mentioning that in this respect, RF might be the algorithm of choice because it is insensitive to tuning especially to the way of tuning (spatial/non spatial) and shows a high performance.  
  
16) Page 23, line 445-446: Please explain: What are the consequences if predictors that profit from autocorrelation are chosen.  
  
17) Page 24 line 465. I think that spatial cross validation is not only recommended but necessary if we have strong autocorrelation and our sampling locations are not evenly distributed  
  
18) Page 24 line 465. It seems like a reference is missing here.  
  
19) Page 25, line 516:521: That's dangerous. We know that it is not acceptable to judge spatial models by a random CV. Please make clear that the high AUROC is not meaningful.   
  
20) page 26 lines 533-536: What does it mean? You cannot judge models by their performance metrics? The meaning does not become clear here.  
  
22) page 27 line 580: 0.05 does not seem to be much of a difference. Is it even significant? Anyway, I suggest leaving this out because your aim was not to improve this prediction by introducing new predictors (at least you never mentioned that this is your aim) but to study the sensitivity of cross validation and hyperparameter tuning so your conclusions should strongly focus on this.  
  
23) Page 27 line 576: " all machine learning models outperformed parametric models in terms of predictive accuracy with RF and GAM showing the best results." That's a contradiction. Earlier on you classified GAM as parametric model.  
  
24) As you submit to ecological modelling, please conclude what is the relevance of your study for spatial modelling in an ecological context.  
  
  
-----------------------------------  
  
  
TABLES AND FIGURES  
  
1) Table 1: Include the default values that you used for the non-tuning models  
  
2) Table 2 is redundant. All information are presented in a more intuitive way in fig.6. Could you leave table 2 out?  
  
3) Figure 3: explain the abbreviations RepCV and SpRepCv in the figure caption. Change the color explanation so that it becomes clear which color is training and which one is test, e.g.  "Blue dots represent the respective training samples and yellow dots represent the respective test samples"  
  
4) figure 5: Please change the label of the x axis so that it's not upside down