Ecomod-18-226R1: Feedback on reviewer#2 comments

Dear Dr Wang,

thank you very much for handling our submission Ecomod-18-226R1. In the first round of reviews we have addressed each of the reviewers' comments and recorded this in a point-by-point reply. We are also thankful for the valuable suggestions which improved substantially the manuscript. We went to great lengths to satisfy both reviewers. The two most substantial changes were the implementation of a spatial tuning for the GAM and the change of the error measure from AUROC to the Brier score. While reviewer #1 was very encouraging and even suggested to accept our manuscript, reviewer #2 expressed a positive overall assessment but remained unsatisfied with respect to several issues. In the previous round of revisions, we had rewritten major parts of our manuscript to address his/her suggestions. At this point, we are in the difficult position to decide how the remaining issues should best be addressed. Therefore, we would like to kindly ask you for your guidance regarding these conflicting issues. In the following, we present the main issues and our thoughts on how and if these can be addressed. Our hope is to reach an agreement with your help, and we are more than happy to make any necessary changes. Your assistance in dealing with these issues is much appreciated.

Point 1: Using the CTF

Reviewer #2 suggests using the idea of the "common task framework" (CTF) by Wilke et al. (2017) (http://www.statisticsviews.com/details/feature/10511089/A-Common-Task-Framework-CTF-for-Objective-Comparison-of-Spatial-Prediction-Metho.html) that Heaton et al. (2017) used to compare algorithms. In our view, the CTF is meant to be both a conceptual framework and a Web-based computing environment for the assessment of machine-learning algorithms on a variety of data sets. We generally agree that the use of multiple datasets would allow more generalizable conclusions, and we welcome Wilke et al.'s contribution that underlines the necessity for standardized assessments. Nevertheless, based on our thorough assessment the CTF's computational framework in its present form is unsuitable for this purpose for a number of reasons:

1. The CTF only contains one dataset (https://hpc.niasra.uow.edu.au/ctf) with three levels of missing data (10%, 30% 50%).
2. It requires the writing of a new wrapper function that accepts training and test datasets. In other words, it does not support cross-validation – not even non-spatial cross-validation – which is a commonly used statistical estimation procedure for the assessment of model performance (Hastie et al. (2001), James et al. (2013)). Reasons for preferring cross-validation over test-set estimation of model performance are well-documented in the literature. In the manuscript we estimate the performance using cross-validation via the *mlr* package, a very well-established, flexible and open-source implementation that ensures reproducibility of our results.
3. The CTF, based on the documentation we have reviewed, does not conveniently support the use of an 'inner' cross-validation on the training set for hyperparameter tuning, which is at the center of our contribution. The *mlr* implementation chosen by us does support convenient tuning and can be regarded as a 'best practice' solution in our view.
4. The CTF runs on R version 3.3.1 from 2016 on a machine with 24 GB RAM. This hardware may be sufficient for a single prediction task of an already tuned model but not for executing spatial CV including hyperparameter tuning in every fold as used in our study (see comment below related to runtime performance).While older R code is usually forward compatible with

newer releases of R and its packages (fingers crossed), we have difficulties adjusting to the idea of trying to run an ecosystem of newer R packages on an older R version for which they weren't written.

5. The description link to the only dataset available in the CTF is broken: http://disc.sci.gsfc.nasa.gov/datareleases/First_CO2_data_from_OCO-2. Overall, it seems that the CTF Web site is not very well maintained.

With respect to all the mentioned points, we think that the CTF in its current state does not add further value to this work.

Reviewer #2 interprets the wording "methodology" to be linked to the "description/development of new methods". In our view "methodology" includes comparing methods/models and analyzing their differences to propose a 'best practice' for hyperparameter tuning.

Regarding the runtime performance of our analysis, we would like to clarify a possible misunderstanding that seems to be affecting our interaction with Reviewer #2. Cross-validation requires repeated model fitting (e.g. 500 fitted models in 100-repeated 5-fold cross-validation), and the use of an inner cross-validation for hyperparameter optimization further increases the computational cost by an additional (large) factor, depending on the particular settings. This was perhaps overlooked by Reviewer #2, who was focused on the test-set estimation procedure implemented in the CTF (see our critical comment above). The use of dedicated high-performance computing resources as available in our department is therefore necessary. We thank the reviewer for pointing us to AWS, which is not necessarily due to the IT infrastructure we have access to.
(four servers, each equipped with 48 cores and 200 GB RAM; overall runtime several days if all cores are used in parallel).

Point 2: Using multiple datasets

Reviewer #2 suggests using multiple datasets. While we agree that multiple datasets enhance generalization, there are multiple practical as well as theoretical issues that come with this:

- We do not claim that the numerical results can be generalized across datasets. We focus on comparing resampling methods (spatial/non-spatial) including hyperparameter tuning on a typical ecological dataset, and how to retrieve a bias-reduced performance estimate in the presence of spatial autocorrelation. We believe that future studies adapting the approach presented in this work will help with finding general patters, e.g. regarding optimal hyperparameter estimates.
- Taking data sets out of their original application context can lead to misleading results as it is hard to identify not just data sets but research questions that fit the exact model type used here. Other data sets might, for example, lead to additional challenges due to multiple levels of grouping, presence of outliers or missing data, all of which would have to be documented in this study. In the machine-learning community the UCI repository (https://archive.ics.uci.edu/ml/index.php) of data sets is frequently used for

performance assessments; it includes one spatial data set, ‘satellite’, which is a remote-sensing data set that is used completely out of context and with no practical relevance. We would like to avoid this type of situation.

- How many datasets should be used? Two or three are maybe better than one but this would still hardly provide general results. A sample of data sets from a ‘population’ of ecological dataset is hard (if not impossible) to obtain, considering also the previous remark. Conversely, benchmarking results obtained by different authors on different data sets using similar methodology may distill into a clearer picture as to which algorithms show superior performances more consistently (which was not the objective of our study).
- If multiple datasets were used, the corresponding result tables will include even more performance results. Stating the fact that we focus on comparing resampling methods, we already have around 30 performance values (5 resampling types times 6 models) for one dataset. Including multiple datasets would multiply this number and make the results and the study confusing for the reader.
- In our perspective, a simulated dataset would not add any additional value to this study in its current state. Additionally, it would again bring up the discussion about “too many results” that was already mentioned in the point discussing multiple datasets.

Overall, there are three major “players” in this study: The algorithms, the datasets and the resampling strategies. Our focus was to compare the resampling strategies and its effect on hyperparameter tuning. We could have only used one model and one dataset to illustrate our point. Adding more models has the added value of an additional model comparison information while still keeping runtime acceptable. Of course, adding more datasets would increase generalization capabilities of this study but as this is not the major focus of this work, the implementation/cost ratio does not match for this point. We hope that the Editor and Reviewer find this view acceptable, and we would appreciate the Editor’s guidance on this issue.

Point 3: Spatio-temporal

Reviewer #2 suggests on using spatio-temporal models because the dataset has a temporal aspect (the response was collected over a period of four years).
The response variable Diplodia sapinea can be present in the area without causing an infection or disease – multiple other factors need to apply to cause a disease in a plot. This makes the temporal aspect a minor one as the spread of the species itself over time does not directly cause infections. Also, if only specific spatial areas would have been sampled each year, the temporal aspect would play a more prominent role in the dataset characteristic as then some observations could introduce a temporal sampling bias.
However, as shown in Figure 1, sample sites were not revisited each year but in fact each year new observations were obtained from the whole study area.
Please also note that we have included the acquisition year of the response as a predictor.
We apologize for not making this spatio-temporal aspect clearer in previous versions of the manuscript. Incorporating all these facts, we hope that the reviewers and editors find a simplification of the dataset characteristic from spatio-temporal to spatial acceptable for this work.

In this study we are interested in modeling disease potential as a function of spatial environmental variables. This approach can be compared to landslide susceptibility modeling where one does not consider the antecedent rainfall conditions (which are very often the cause triggering landslides) but only static topographic and other environmental variables as predisposing factors.

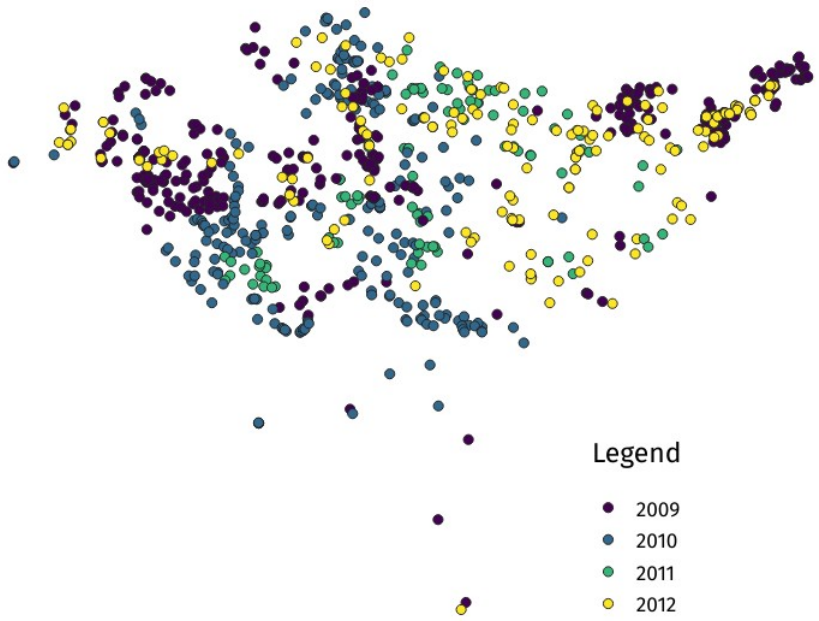

Figure 1: Fig. 1: Spatio-temporal distribution of the sampling of the response variable

Point 4: Spatial GLM

We are well aware that a GLMM is just an extension of a GLM with a random component that can either consist of a random effect or a spatial autocorrelation structure.
Reviewer #2 does not agree with our assumption of an equal performance between parametric models with and without a spatial autocorrelation structure.

To confirm our assumption empirically, one would need to estimate the spatial autocorrelation structure for each model of a CV (i.e. for 500 models in our case) and train a GLMM instead of a GLM and compare performances. In our view, estimating an autocorrelation structure on the full dataset only and using this one in all models of the CV would introduce a bias, therefore this simplification would not be an option. Strictly, the same would have to be done for the GAM. All of this would go beyond the scope of this work.
We are not aware of any research that has shown that a GLMM including a spatial autocorrelation structure increases performance in situations comparable to the present one (i.e. not interpolation). We also could not find any evidence in the references of the first review of reviewer #2.

Summary

We would like to emphasize that we support initiatives such as the CTF proposed by reviewer #2. However, we don't think the current implementation is mature enough to be used in comparison studies like this in which runtime plays a role.
The CTF could integrate more datasets, possible sources are listed below:

- R package "spData": https://cran.r-project.org/web/packages/spData/index.html.
- https://www.gbif.org/
- https://www.givd.info/

We are open to use mixed models if there is convincing evidence that incorporating spatial autocorrelation structures increases performance.

We hope that, based on our comments and explanations, the editor can give us guidance on how we should incorporate Reviewer #2's recommendations into our work, considering that both reviews are generally positive. We hope that this conflicting situation can be solved and apologize for the inconvenience caused by this. We thank both reviewers for their valuable time and input which certainly improved the quality of this manuscript.

References

Hastie, T., Friedman, J., & Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-21606-5

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., & Nauss, T. (2018). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software*, *101*, 1–9. https://doi.org/10.1016/j.envsoft.2017.12.001