Reviewer #2 comments: Revision #2 of "Crucial but often neglected: The important role of spatial autocorrelation in hyperparameter tuning and predictive performance of machine-learning algorithms for spatial data."

Dear Reviewer #2,

we greatly appreciate your detailed comments on the manuscript. Changing from AUROC to Brier score was a major improvement of the study. In addition, we incorporated most of the smaller changes you suggested.

However, we kindly disagree with certain points of your second review. In the following we outline our thoughts in detail. We addressed all these points also in the manuscript in a more brief version. It is important to discuss them and we are glad you brought them up. However, they are either not always contributing to the overall goal of this work or are still in scientific discussion. See below for the details.

In my previous review, my first major comment was:

• By using only a single data set, it is not clear how generalizable the results are. It seems that more than one data example should be used. I suggest framing the comparison using a common task framework (sensu Wikle et al. 2017). Heaton et al. (2017) provides an excellent example applying the common task framework to evaluate methods for spatial prediction.

The authors suggest that the focus of their work "is on the methodology (spatial tuning of hyperparameters, influence of spatial autocorrelation) rather than generalization across datasets" and provide many straw man arguments as to why they can't make their study rigorous and complete (e.g., "there is only one dataset available" and "the processing power is limited (6 cores, 24 GB RAM)") My first comment is that if the paper is really focused on methodology, why not just use, in addition to a real data example, a simulation study where "truth" is known? For example, it makes little or no sense to talk about "unbiased performance estimation" when you don't have a gold standard (i.e., when truth is known).

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.

The authors suggest that the focus of their work "is on the methodology (spatial tuning of hyperparameters, influence of spatial autocorrelation) rather than generalization across datasets" and provide many straw man arguments as to why they can't make their study rigorous and complete (e.g., "there is only one dataset available" and "the processing power is limited (6 cores, 24 GB RAM)") My first comment is that if the paper is really focused on methodology, why not just use, in addition to a real data example, a simulation study where "truth" is known? For example, it makes little or no sense to talk about "unbiased performance estimation" when you don't have a gold standard (i.e., when truth is known).

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.

I think it is non-sense to say that the focus of this paper is on methodology (i.e., the description/development of novel methods), when clearly the authors are trying to evaluate which "methods" are best.

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

Second, the reasons the authors list are trivial. For example, a modern laptop computer has 6 cores and 24 GB or RAM. Sure it may be a slight inconvenience to do a rigorous study, but surely it doesn't take weeks to run (on a laptop) and if the author's wanted they could use higher performance computing (e.g., a desktop or something like AWS; see http://www.louisaslett.com/RStudio_AMI/ for an easily accessible solution). Also, I don't think saying "there is only one dataset available" is helpful. I am sure the authors could find another data set (e.g., take a look here https://datadryad.org/)

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. (A HPC cluster with 6 compute nodes, each equipped with 32-48 CPUs and 200 GB RAM; overall runtime several days if all cores are used in parallel). This is due to the extensive hyperparameter tuning that is conducted.

In my previous review, my third major comment was

• I suspect that the data used in this paper (e.g., Figure 1) are spatio-temporal (i.e., not all collected at the same time). Although spatial prediction is widely used in ecology, it seems the trend is towards making spatio-temporal predictions using statistical models and machine learning algorithms (e.g., Hefley et al. 2017).

The authors confirm that the data are spatio-temporal and collected over a 4 year time period. The authors response is non-sense (e.g., “all observations are unique in space, meaning that there is no spatio-temporal overlap.” and “due to the long-term average characteristic of most variables (e.g. temperature, precipitation, etc), the temporal aspect of the response variable becomes less important”). If the issue of spatial autocorrelation is such a huge concern when tuning hyperparameters (which is the main impetus for the authors work and even show up in the title!), then why wouldn’t temporal autocorrelation matter? I think the authors need to address the spatio-temporal aspect of their data example or just get rid of the data example and use simulated data in a spatial-only setting. As written, this paper is misleading.

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.

Finally, on pg. 27 lines 540 - 543: The authors suggest that the predictive accuracy of parametric models (e.g., GLMs) with and without spatial autocorrelation structures is perhaps the same and that little research has been done on this topic. This statement is blatantly wrong. In most cases, a parametric model without an effect that accounts for spatial autocorrelation (e.g., a spatial random effect) is a special case of the spatial model (e.g., a spatial generalized linear mixed model is just a GLM with a spatial random effect). In almost all non-pathological cases the spatial effect will increase the predictive accuracy. I suggest that the authors look at some of the formal statistical references in my first review. Overall, the quality of the entire paper could be improved if the authors paid some regard to the primary literature in statistics and machine learning rather than rely so heavily on the secondary (and often incorrect) ecological literature.

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. You disagree 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.