Angers, September 4th, 2020

Objet: Response to reviewers

Dear Editor,

Thank you for your message. We would also like to express our warm thanks to the reviewers for their thorough and helpful assessments. After a careful reading of the reviewers' comments, we have done our best to take their comments and suggestions into account and we hope that the new version of our manuscript has been improved in terms of quality.

Reviewer 1

The manuscript addresses the use of citizen science data to evaluate SDMs accuracy as independent evaluation dataset. This is a relevant and timely question given the increasing abundance of this kind of data, which can undoubtedly be of great help in the field of ecological modelling. The manuscript also considers some of the main limitations of this occurrence data (observer expertise, heterogeneous sampling effort…) for both calibration and evaluation datasets. However, I found that some methodological issues — mainly related to this sampling effort bias — were poorly explained given their relevance for manuscript results and conclusions.

Major comments:

- In section "Background data and pseudo-absence selection", authors explain three different strategies used for generating pseudo-absences for model calibration. In case of s3 strategy, it is supposed to deal with sampling effort bias in three ways: accessibility bias, attractiveness bias, and observer sampling effort. However, I found no specific explanation of how this bias was considered neither in the main text nor in the supplementary material. Dealing with bias is a particularly sensitive issue, since it is often highly difficult to know precisely to what extent an area has been oversampled. Without a detailed explanation of the bias treatment it is difficult to discuss whether bias management are properly applied or instead generate new bias sources. So, further clarifications in the main text or supplementary material about this bias treatment are required.

Authors: we have added an appendix (see Appendix 3) "Definition and modelling of sampling effort for SDM at regional extent"

- Overall, I liked the manuscript focus and approach, however I find that the methodological scope of the introduction blurs a little in the discussion. I miss in the discussion any mention of the shortcomings of using citizen science data directly for model evaluation, since actually in this study CS.0 data quality somehow forced to complement the database with new data of volunteers and professionals during 2018-2019. In addition, I think that the manuscript could benefit from placing the results (obtained in a local example) in a more general context and discuss the possibilities and limitations of using these citizen science data to validate models of other species on larger geographical scales.

Authors: we have added some lines in the discussion (395 to 416)

Minor comments:

Line 33: Specify crossvalidation with internal evalutation dataset (in contrast to external evaluation).
Authors: ok
Line 34: Here and in other manuscript sections appear "overassessment". Likely "overestimation of model accuracy" is a more accurate expression.
Authors: ok
Line 68: Likely the paper audience will understand what is opportunistic presence-only data. However, it could help to briefly define opportunistic presence-only data from citizen science vs detection-non detection citizen science data.
Authors: ok line 71 – "Presence-only data come from different source databases reduced to simple species presence records and mostly collected in an unstandardized way by volunteers. In contrast to presence-absence data, they are abundant but have poor quality, few metadata and come from different sources (Robinson et al., 2020)."
Line 81: According to the beginning of the sentence, it seems that it lacks a reference before the full stop.
Authors: ok, reference added.
Line 81-83: This sentences seems misplaced here. Rewrite.
Authors: ok. Sentence has been placed in methodology.
Lines 136: substitute "amphibians" by "some amphibians species".
Authors: ok
Lines 141-143: Specify here the list of analyzed amphibian species. In addition, the paragraph could be rewritten in order to clarify that opportunistic citizen science was used as calibration database both when using opportunistic citizen as evaluation data and detection-nondetection as evaluation data.
Authors: ok. Line 141 – "We studied habitat suitability of 9 amphibian species: Bufo spinosus, Hyla arborea, Rana dalmatina, Rana temporaria, Triturus cristatus, Triturus marmoratus, Lissotriton helveticus, Salamandra Salamandra, and Pelodytes punctatus. Two types of amphibian data were used: (1) opportunistic data from a citizen database with presence-only records for model calibration and internal evaluation (2) standardized detection-nondetection data from a citizen science programme and complementary field work for external evaluations. A more detailed description of the data sets and complementation strategies is available in Appendix 1."
Line 146: Please clarify some details of the citizen database: access web, citizen science program name, and/or responsible institution.
Authors: ok – in text and Appendix 1 – line 149 – "We accessed presence-only occurrences from a regional database for the period 2013-2019. 86% of the dataset was collected by citizens and recorded online (website or associated mobile application) and 14% by various professional organisations involved in nature protection. All data were compiled for the regional Atlas of amphibians by a French non-governmental organisation (French BirdLife partner - LPO). See Appendix 1 Table 1 for data sources."
Line 159: Same as line 146.
Authors: ok – in text and Appendix 1 – line 164 – "For external validation, we firstly extracted detection-nondetection amphibian data for the period 2013-2019 collected as part of a citizen science program called "Un Dragon dans mon Jardin" (Appendix 1 section 1.2.)."
Line 160: "one year". Do authors mean "each year"? In addition, specify what are good weather conditions.
Authors: ok – "We retained 576 sites which were monitored at least 3 times between February and June during at least one year" / complements in Appendix 1
Line 163-164: Rewrite. Briefly mention at the beginning the CS.0 shortcomings which made it necessary to complement the database with new fieldwork.
Authors: ok. Line 169 - "Some large areas of the region were not sampled due to lack of observers so that data were clustered near cities, with spatial autocorrelation. Therefore, with help from volunteers, we completed this dataset with some additional fieldwork and applied filters."

Lines 168-171: Collection approaches are not clear enough. 108 sites were proposed for monitoring by volunteers but finally, only 75 was monitored by them. Rewrite this section to clarify or consider to remove lines 168-169.

Authors: ok – sentence has been removed

Line 175: Why authors specifically use 5% value? How do they deduce that urbanized areas and croplands are exactly 5% less sampled than the rest of areas?

Authors: There was no data in these land cover types. 10% represents their proportion in the region. With random sampling, 5% of the data would have fallen in highly intensive croplands and 5% in urban areas. Our approach is debatable but we did not find any other satisfactory solution.

Lines 197-200: Please explain more this stratified sampling approach.

Authors: ok. We add one sentence

Lines 206-207: Specify climatic data source. 1950-2000 of Hijmans et al. 2005 correspond to Worlclim V1 database. This is an old version, so if authors have recently download the climatic data, refer to current version (v. 2.1 period 1970-2000). In addition, better refer to spatial resolution in km2 (in this case 2.5 arc-min ~ 5 km2 at equator).

Authors: This version 2.1 was released in January 2020. We made the analyses in 2019 with the v1.

Line 210: This paragraph specifies a 2.5 arc-min (~ 5 km2) spatial resolution, however fig 2 from Appendix 2 show the correlation circle of variables with 500m. How do authors get this last value? Correlation circle was obtained before or after climatic values extraction of species occurrences?

Authors: We converted the source raster (with 5km resolution) to a raster of 500m resolution for homogenisation with other variables. This manipulation is needed to stack all environmental variables together in R and make the manipulation easier. The PCA was implemented with 12 Worldclim layers to obtain 2 regional raster layers corresponding to the 2 first axes of the PCA. The results are the same with 5km resolution rasters. Variable transformations were done before extraction of species occurrence values.

Lines 255-258: Clarify whether ensemble process will affect Random Forest and GAM models separately or whether it will join both algorithm predictions.

Authors: we have tried to clarify in the text. Line 271 – "Finally, we conducted ensemble modelling by calculating the median value of (1) all individual maps generated by GAM and Random forest (i.e. 500 maps/algorithm) (Thuiller et al., 2009) to compare internal versus external evaluation for each species. Secondly, we also calculated median values from ensemble maps calibrated with 100% of presence-only data to compare different external evaluation sets (i.e. 10 maps/algorithm)."

Line 268: Substitute "is the most accurate" by "is considered to be".

Authors: ok

Line 269: I would recommend to remove this sentence.

Authors: ok, sentence deleted

Lines 296-297: If I have properly understood, different models are s1, s2 and s3 and are being selected depending on AUC values (with each different evaluation dataset –internal and external-). Please rewrite and clarify.

Authors: yes. We have tried to clarify the sentence. Line 295 – "For each species, the median AUC was higher with internal validation than external validation for all three pseudo-absence selection strategies (s1, s2 and s3), both for GAM and Random Forest (see Figure 2 and Appendix 4)"

Lines 308-309: Correct me if I'm wrong, but it seems that *Lissotriton helveticus* and *Hyla arborea* had considerably lower accuracy values for PRO dataset than STRAT_CS and STRAT_ALL.

Authors: there a mistake in the sentence. We have clarified. "The models selected (s2 or s3) were similar for most species whether using stratified data from volunteers' only (STRAT_CS) or stratified data with added professional observations (STRAT_ALL), or professional data only (PRO). See Table 2 and Appendix 4. We excluded the s1 model from the comparison because this model is never selected, either with internal or external evaluation."

Line 335: Phillips et al. 2009 also found increases in AUC when considering sampling effort bias. This reference may help for discussing.

Authors: Thank you. Phillips et al. 2009 used external and independent dataset for external evaluation. Our problem line 335 is more about the selection of the model corrected for sampling effort with internal validation. It's not the case for other species, especially because internal evaluation dataset is biased in the same way as the calibration set. For these species, sampling bias correction is not adapted. They are mostly forest specialists and forest is rare in our study region. We added a sentence line 379. "In addition, filtered evaluation dataset showed coherent results according to Phillips et al. 2009. Indeed, choosing pseudo-absence data with the same bias as occurrence data improved model performance."

Lines 356-357: It's true that using independent validation data avoids to split the occurrence dataset, which lead to reductions in the number of records used for model calibration. However, it should be considered that when occurrence datasets is too scarce it is also possible to split the dataset between calibration and test set, and then make model predictions using all records (Franklin 2010).

Authors: Sorry, we don't understand this comment. Are you talking about substitution procedure?

Figure 3: If possible, it could be useful to add a little map of France to easily locate the study area. Some figures of supplementary material specify that s3 pseudo-absence generation strategy also excluded known presence points, however Fig.3 does not show this specification. Please homogenize captions. Finally, how did authors project models with 500 m2 resolution if climatic database had 2.5 arc-min? Authors should describe projection procedure (and climatic downscaling in such case) in method section or supplementary material.

Authors: we added a map of France and homogenized captions.

Supplementary Material: Some figure and table captions are under-explained (Fig 3 Appendix 1, Fig 2 Appendix 2). Fig 1 Appendix 2 legend is not readable. Appendix 2 bibliography title should be in English. In addition, it would be better to reorder appendix figures and tables according to manuscript appearance order and to specify table and figure number instead of just appendix name (e.g. line 210 Figure 2 Appendix 2 instead of Appendix 2).

Authors: ok – we complete Appendix references and ordered Figures.

 # Reviewer 2

The paper entitled «*How citizen science could improve Species Distribution Models and their independent assessment for conservation*» (M Florence, J Baudry, G Pain, M Sineau and J Pithon) tests various ways of using data originated from a citizen science project in a French region to generate spatial distribution models (SDM). The paper, which is mostly methodological, is well structured, understandable and focused, and the bibliography is quite updated. My main concert, which does not invalidate the approach nor the whole work, is they way data have been selected, filtered, or how some thresholds have been decided. Here are my suggestions/comments (that include the previous comment), that will hopefully help you improve the manuscript:

1. I do not see how the authors develop the second part of the title «*.... and their independent assessment for conservation*». The species used as biological models belong to an endangered taxonomic group, but the paper is clearly focused on the «HOW citizen science....» part. SDM and maps are used for conservation, but they also have other applications. The authors could have used non-endangered species in the same way. In the discussion there is not much about this second half as well, so if they do not work on this (probably in the discussion), I would just take it out.

Authors: ok – we have deleted “for conservation” in the title and completed some information about our project in the last paragraph line 417. In fact, our research is part of a wider regional project concerned with amphibian conservation in human-modified landscapes. In this project, we have worked to strengthen the links between researchers, non-academic structures and volunteers to strengthen a conservation objective and to improve the applicability of SDMs for conservation purposes. For this reason, we maintain our discussion of the impacts of using different SDM approaches on possible conservation action.

2. Where does the «Opportunistic presence data» data set comes from? This information does not appear neither in the main text nor in the Appendix. The only reference is «from a regional database for the period 2013-2019» in L147.
Authors: ok – we clarify in text and Appendix 1. Line 149 – “We accessed presence-only occurrences from a regional database for the period 2013-2019. 86% of the dataset was collected by citizens and recorded online (website or associated mobile application) and 14% by various professional organisations involved in nature protection. All data were compiled for the regional Atlas of amphibians by a French non-governmental organisation (French BirdLife partner - LPO). See Appendix 1 Table 1 for data sources.”

3. The data set for «Standardized detection-nondetection data (external validation data-sets)» is the same as the Opportunistic? The explanation is exactly the same: «we firstly extracted detection-nondetection amphibian data from a regional citizen science database for the period 2013-2019». I understand it should not, and that this comes from «Un dragon dans mon jardin», but as it’s explained now is quite confusing.
Authors: ok – we clarify in text and appendix 1 – “For external validation, we firstly extracted detection-nondetection amphibian data for the period 2013-2019 collected as part of a citizen science program called “Un Dragon dans mon Jardin” (Appendix 1 section 1.2.).”

4. When in the paper the authors cite the supplementary material, the reference is only «Appendix X». This is clearly not enough, because the 3 files have various figures and/or tables. The table or the Figure in the Appendix should also be mentione to make the reading easier.
Authors: ok – we complete Appendix references

5. Data from CS.0 «Un dragon dans mon jardin». The data that the web page shows (https://www.undragon.org) does not seem to match with the data you say you use. For example, the web page shows only 12 observations of *Bufo spinosus* in Pays de Loire, whereas in your Table 2 of Appendix 1, there are 79 cells with detection of the species. It may be the number of years (I have not found the way to filter by year), quality-related filters, or any other reason, but this difference (which may occur with other species as well) should be explained.
Authors: Yes sorry. We just cite the website of the general program. All data sources are now detailed in Appendix 1 Table 1.

6. L182-193. Criteria 2 to set the threshold to validate non-detection as absence data. The way of setting the thresholds is relevant, because it excludes or includes presence data. It should be very clearly specified, and it seems it's not. As explained it seems it’s been manually/grahically-driven carried out. Has it? An alternative, although it represents a whole change of the statistical approach, would be using hierarchical models, which also take into account detectability. It’s not that I think that you MUST use this approach, but if you do not, the way thresholds have been set must be clearly explained.
Authors: We have clarified the use of Boissinot’s 2008 work in Appendix 3. He defined minimum sampling effort to get 0.95% focal species detection probability according to the method used (calls

survey, active search or fishing). He studied the same species as us in a similar ecological context. We use these results to fix the threshold for expert, intermediate and beginner observers. However, it's true than we could have use an occupancy model to calculate detection probability according to our observer level groups. We judged that the knowledge of species detection available in the literature was sufficient.

7. Background data and pseudo-absence selection (L240-L253). Is this only for the Opportunistic data-set? It's not specified, but I understand that the data-set from the «Un dragon dans mon jardin» and the data coming from PRO and VOL has also real absences, right?
Authors: Yes, we try to clarify the difference between data sets line 143 – "Two types of amphibian data were used: (1) opportunistic data from a citizen database with presence-only records for model calibration and internal evaluation (2) standardized detection-nondetection data from a citizen science programme and complementary field work for external evaluations. A more detailed description of the data sets and complementation strategies is available in Appendix 1."

8. There is no much detail on how "s3: random pseudo-absence selection constrained to take sampling effort into account." has been done. This should be clearly specified, at lease in the Supplementary.
Authors: we have added an appendix (Appendix 3) "Definition and modelling of sampling effort for SDM at regional extent"

9. I think that the section «Involved stakeholders and citizens in the research process» must be shortened and changed. In my opinion, most of the paragraph is too much generic and common for most CS projects. I would highlight, though, this sentence: «*Involving citizens at different stage of the mapping process may make action easier to implement, through both better shared knowledge and stronger personal involvement.*». This sentence fits well with the whole work, and making suggestions **how** citizens may be involved in **which stage** will be a value in this paper, instead (or in addition to) these more generic sentences that do not really add much to the paper.
Authors: We have shortened and changed this section. Lines 418-429

<u>Some minor comments:</u>
- Some axis titles do not start with capitals. Please correct
- Figure 1, Table 3 (in Appendix 1) Missing the description of the acronym «ABS» in the caption
- Figure 1 Appendix 1, missing units of the Y-axis
- L264: indicate that (1) is the internal model validation
-L303: change «specie» by «species»
- L412: change «Bibliographie» by «Bibliography»
Authors: ok