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.
- 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.

Minor comments:

Line 33: Specify crossvalidation with internal evalutation dataset (in contrast to external evaluation).

Line 34: Here and in other manuscript sections appear "overassessment". Likely "overestimation of model accuracy" is a more accurate expression.

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.

Line 81: According to the beginning of the sentence, it seems that it lacks a reference before the full stop.

Line 81-83: This sentences seems misplaced here. Rewrite.

Lines 136: substitute "amphibians" by "some amphibians species".

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.

Line 146: Please clarify some details of the citizen database: access web, citizen science program name, and/or responsible institution.

Line 159: Same as line 146.

Line 160: "one year". Do authors mean "each year"? In addition, specify what are good weather conditions.

Line 163-164: Rewrite. Briefly mention at the beginning the CS.0 shortcomings which made it necessary to complement the database with new fieldwork.

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.

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?

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

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 km^2 (in this case 2.5 arc-min ~ 5 km^2 at equator).

Line 210: This paragraph specifies a 2.5 arc-min (~ 5 km^2) 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?

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

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

Line 269: I would recommend to remove this sentence.

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.

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.

Line 319: Which particular criticism?

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

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).

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 m^2 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.

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).

References:

Franklin, J. 2010. Mapping species distributions. Spatial inference and prediction. - Cambridge University Press.

Phillips, S. J. et al. 2009. Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. - Ecol. Appl. 19: 181–197.