# Choosing presence-only species distribution models

Boris Leroy

Received: 19 January 2022 | Revised: 22 August 2022 | Accepted: 11 September 2022

DOI: 10.1111/jbi.14505

Journal of Biogeography WILEY

COMMENTARY

# Choosing presence-only species distribution models

**Boris Leroy**

Unité 8067 Biologie des Organismes et Ecosystèmes Aquatiques (BOREA), Muséum national d'Histoire naturelle, Sorbonne Université, Université de Caen Normandie, CNRS, IRD, Université des Antilles, Paris, France

**Correspondence**
Boris Leroy, Unité 8067 Biologie des Organismes et Ecosystèmes Aquatiques (BOREA), Muséum national d'Histoire naturelle, Sorbonne Université, Université de Caen Normandie, CNRS, IRD, Université des Antilles, Paris, France.
Email: leroy.boris@gmail.com

**Funding information**
Permanent position from French government

**Handling Editor:** Daniel Chapman

Over the past two decades, species distribution models (SDMs) have become one of the most popular modelling tools in biogeographical studies. SDMs try to quantify the relationship between a taxon and its environment, for example, to predict its geographical distribution, to assess potential impacts of climate or land use change, or to explore biogeographical hypotheses. In practice, SDMs generally correlate species distribution data in the form of spatially explicit presences and absences, to environmental predictors, such as climatic variables. In cases where presences and absences are difficult to obtain in quantity and quality—that is, for the majority of biodiversity—it is possible to use SDMs with presence data alone. These are dedicated approaches requiring the generation of additional data points (called 'background points' or 'pseudoabsences'). Overall, the concept of SDMs is simple; however, their implementation is complex because a large number of decisions are required throughout the multiple steps of the process (Figure 1). Each of these decisions must be weighed carefully by the users because they have a strong influence on the outcomes of SDMs and their interpretation. Guidance on how to make these decisions can be found in methodological or pedagogical papers and books (e.g. Elith et al., 2006; Guillera-Arroita et al., 2015; Guisan et al., 2017; Guisan & Thuiller, 2005; Phillips et al., 2006; Thuiller et al., 2009). However, for the majority of these decisions, there is still a high degree of uncertainty, because of shortfalls in our knowledge (see my perception of this degree of uncertainty in Figure 1). This uncertainty often leads to either making arbitrary decisions or costly sensitivity analyses when preparing SDMs. Furthermore, the profusion of methodological studies makes it easy for users (especially new users) to either miss guidance or caveats relevant to their study, or to lack the ability to understand them.

Two main issues in SDM implementation require further guidance: uncertainty in decision-making and inaccessibility of guidelines. Uncertainty can be addressed by studies comprehensively investigating a specific methodological issue, providing established guidelines for decision-making. Inaccessibility can be addressed by studies synthesising the methodological progress with sufficient pedagogy to propagate good practices in the field. Here, I appraise a recent study which has combined these two characteristics in such an outstanding way that it should be extremely helpful to both new and experienced users (Valavi et al., 2021). First, Valavi and colleagues comprehensively addressed the choice of modelling techniques in presence-only situations, which has been a prominent issue so far. Second, they detailed all their methodological choices pedagogically, explaining the underlying reasons, and thus providing accessible guidelines throughout the multiple steps of the modelling process. In the following text, I first explain the context of model choice and pinpoint the major progress provided by Valavi and colleagues, and then I explain why, beyond this progress, their study will improve practices in the field. Finally, I conclude with an outlook on the uncertain decisions in SDMs.

## 1 | ADDRESSING THE GAP OF MODEL CHOICE IN PRESENCE-ONLY STUDIES

SDMs can be calibrated with many modelling techniques (e.g. Valavi and colleagues compared 21 techniques), which can be classified

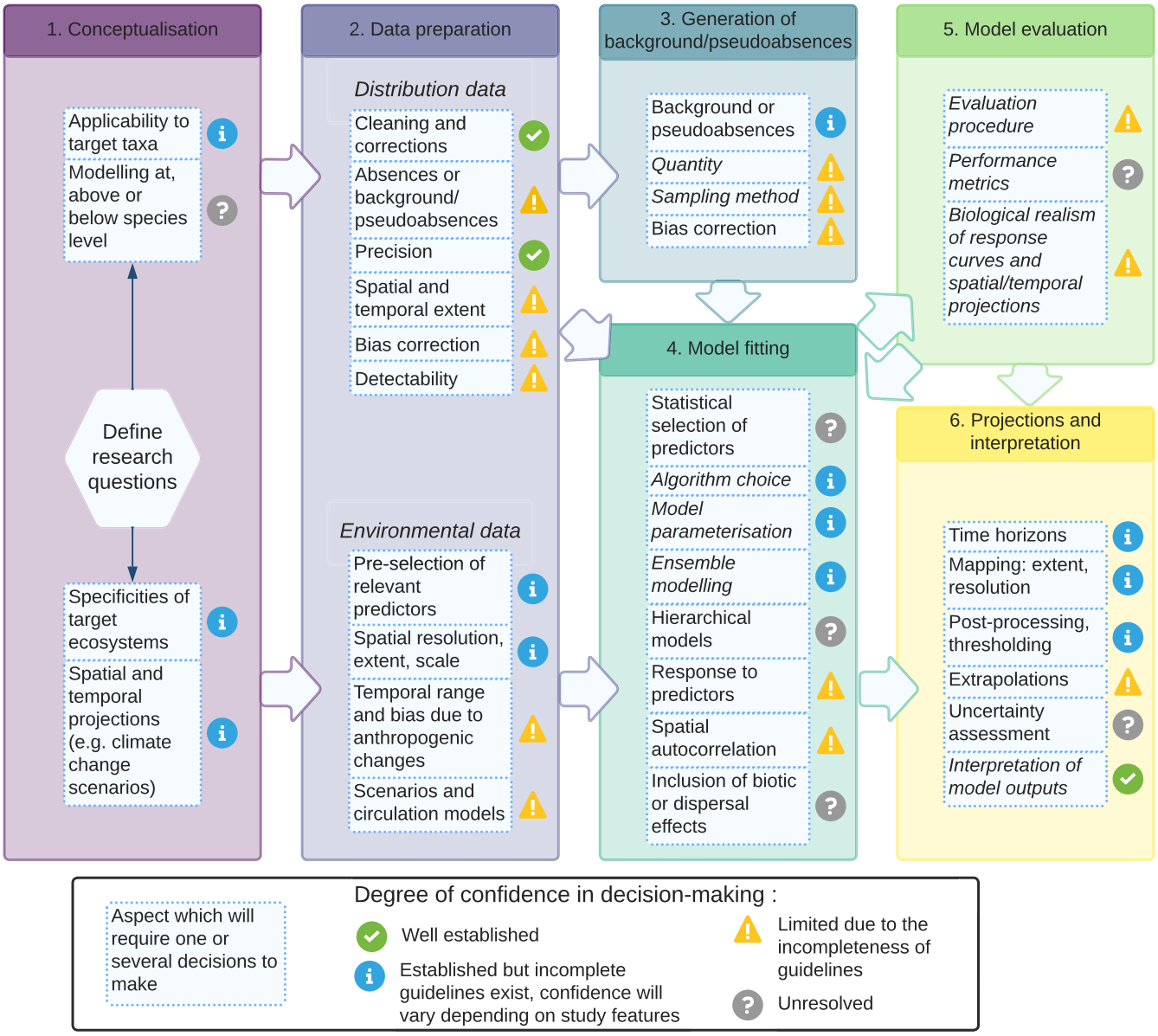


**FIGURE 1** Schematic representation of the major steps of the process of building a species distribution model (SDM) and the multiple decisions that users have to make at all steps. Each decision will impact the outcome of models, and must be carefully chosen by users by applying guidelines from the literature to their study. However, there is still a large degree of uncertainty in our knowledge for best decision-making, as illustrated with the different confidence icons. This figure was designed to illustrate a comprehensive view of the general complexity of building a model, but it is not exhaustive and its content should be expected to vary among studies because of the many idiosyncrasies in SDM studies. Italicised text indicates a decision which was discussed in Valavi et al. (2021). Note that the order of decisions inside each step does not represent a sequential order in which steps should be undertaken

into two broad categories: (1) regression approaches such as generalised linear models and (2) classification-based algorithms and their machine-learning extensions, such as random forests (see Guisan et al., 2017 for a detailed overview). New techniques are continuously being introduced into the field, such as new classification-based algorithms (e.g. XGBoost Chen & Guestrin, 2016), improving classical algorithms such as boosted regression trees (Elith et al., 2008). Conversely, some classical approaches are still subject to progress in their understanding, such as the maximum entropy algorithm (MaxEnt, Elith et al., 2011) which has been recently associated with a range of statistical methods called point process models (Renner et al., 2015). The diversity of possible algorithms for SDMs makes the selection of candidate algorithms a difficult decision, especially in presence-only situations where it is notoriously difficult to estimate model performance reliably. The field of biogeography was sorely lacking a comprehensive and reproducible evaluation of the performance of the currently available algorithms.

To fill this gap, Valavi and colleagues used a massive benchmark dataset composed of 225 species with presence-only records from six different regions of the world. This dataset has several features

which explain why it is suitable to provide recommendations on model performance which can also be generalised to areas beyond this study. First and foremost, it has independent presence-absence data to evaluate the presence-only models for each species. Independent presence–absence data for model evaluation are the most appropriate and arguably the rarest type of evaluation data in presence-only studies. This first feature itself sets this comparative study as a reference in the field for the choice of presence-only modelling techniques, since no other study has been published on this topic with a similar or better evaluation dataset. Second, the environmental conditions and taxonomic groups in this dataset are diverse. This diversity probably covers a large breadth of relationships between species and their environment, and thus the results could be transferred to a wide range of taxa with similar species–environment relationships. Third, the dataset covers a gradient of data in quantity (from five to thousands of records) and quality (biased and unbiased). Hence, the results of Valavi and colleagues help modellers to decide on which techniques are adequate for their study, by pinpointing the techniques which perform best under data-poor or data-rich conditions, and those that perform well in all conditions.

Valavi and colleagues reached conclusive results on the relative performance of the 21 tested modelling techniques. I do not intend to mention all their important results, but rather, I want to illustrate with some examples how the findings are likely to improve future practices. For example, they established how the choice of default settings over carefully tuned models can be extremely detrimental to model performance, which is the case of the machine-learning technique called Random Forests. They showed that Random Forests are the best individual technique with a few well-tuned parameters, and the worst technique under default parameters. Likewise, they showed that the combination of multiple modelling techniques to obtain a consensus prediction, a method called 'ensemble modelling' (Araújo & New, 2007), can be the best performing procedure if based on well-tuned models, whereas it will perform no better than average if based on models with default settings.

## 2 | GUIDANCE FOR IMPLEMENTING SDMS

Yet, methodological progress is only the first of the two reasons why Valavi et al. (2021) are likely to improve SDM practices. The second reason lies in the pedagogy of the paper. They explain these complex methods in a text which is accessible to all researchers who have an ecological background, but not necessarily a strong modelling background. Specifically, one of the main barriers in choosing or tuning modelling techniques lies in user understanding, or the lack thereof. Descriptions of techniques are available in the literature, but are not necessarily formulated in a way which can be readily understood by ecologists, and this can lead to misuse, misinterpretation or avoidance of techniques. Valavi and colleagues made biologically meaningful descriptions of the different techniques and parameterisations used. Besides the description of techniques and the choice of parameters, Valavi and colleagues also thoroughly explained their methodological decisions pertaining to other steps of the process, such as the selection of background points or the evaluation procedure, synthesising existing knowledge in the literature (italicised text, Figure 1) in an accessible prose. Furthermore, Valavi and colleagues provided their code with sufficiently detailed comments to be understood and reproduced. Therefore, they show how easy it is to adopt these new best practices straight away, and they provide new users with the necessary tools to do so. For all these reasons, Valavi et al. (2021) become a reference in the list of guidance papers to use when conducting SDM studies (Araújo et al., 2019; Feng et al., 2019; Guillera-Arroita et al., 2015; Guisan & Thuiller, 2005; Zurell et al., 2020), especially for researchers beginning with SDMs.

## 3 | OUTLOOK

Valavi et al. (2021) will become the long-awaited benchmark for model comparison in presence-only situations. However, in spite of the diversity of species and regions it covers, it still remains restricted to plants and vertebrates of continental environments which is a potential limitation to its generalisability. Thus, an outstanding issue to explore is the limits of this dataset, and whether it can be generalised to a broader scope of organisms and environments, especially when species–environment relationships are likely to be different. For example, it is difficult to foresee whether functional responses for invertebrate taxa would result in similar model performance for plant and vertebrate taxa. Likewise, marine, freshwater or subterranean environments have specific sets of predictor variables whose nature may differ from continental climatic variables, which, in turn, may result in a different ranking of model performance. Assessing the transferability of these results to other types of organisms and environments is, in my opinion, a crucial question to address in the future.

Regarding a related yet broader perspective, the methods of, or constraints on, variable selection and interpretation were not addressed by Valavi and colleagues, probably because they focused on the predictive performance of SDMs. However, the objective of SDMs can also be the understanding of the drivers of species distributions, and the process of identifying and interpreting important variables remains an outstanding issue which needs to be comprehensively addressed. In fact, we can extend this observation to the many modelling decisions for which the literature offers insufficient guidance or currently unresolved knowledge (Figure 1), such as, to name a few that I find important: how to identify the optimal taxonomic resolution at which a niche should be modelled; how to choose between pseudoabsences and background points and what is the optimal sampling strategy in each case; how to best combine multiple model layers, for example, to combine a mechanistic with a correlative model or models calibrated at different spatial scales; how to properly evaluate models accounting for both performance and biological realism; and how to evaluate and communicate comprehensively the uncertainty of SDMs.

Insofar as many of the decisions required to implement SDMs lack established guidance, I urge ecologists to communicate the rationale for their choices more clearly. The reporting of decisions in SDM studies has recently been improved thanks to several papers aimed at standardising the documentation of protocols (Araújo et al., 2019; Feng et al., 2019; Zurell et al., 2020). However, there is still progress to be made in providing the reasons underlying uncertain or debatable choices, and such progress will have multiple positive implications for the field. First, adopting such practices will improve the review process, since the reasons underlying methodological choices are frequently asked for. Second, some of the decisions taken have consequences on the interpretation of model predictions, but this link may remain implicit to the reader unless the authors specify it explicitly. Last, airing rationales and hypotheses supporting choices will generate transparent discussions and debates in the literature, which in turn, will enable collective progress towards reducing uncertainty in SDM implementation.

## ACKNOWLEDGEMENTS

I thank Céline Bellard, Nicolas Dubos, Eric Goberville, Berta Ramiro-Sánchez, Cam Ly Rintz and Coline Royaux for fruitful discussions on the content of this commentary. I am thankful to Damaris Zurell and Mike Dawson for their comments which helped me improve this manuscript. I thank Aldyth Nys for editing the English language. I was funded by my salary as a French public servant. No permit was required to achieve this work.

## CONFLICT OF INTEREST

I declare no conflict of interest.

## DATA AVAILABILITY STATEMENT

No data were used in this commentary.

## ORCID

*Boris Leroy* https://orcid.org/0000-0002-7686-4302

## REFERENCES

Araújo, M. B., Anderson, R. P., Barbosa, A. M., Beale, C. M., Dormann, C. F., Early, R., Garcia, R. A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R. B., Zimmermann, N. E., & Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. *Science Advances*, *5*(1). https://doi.org/10.1126/sciadv.aat4858

Araújo, M. B., & New, M. (2007). Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, *22*(1), 42–47. https://doi.org/10.1016/j.tree.2006.09.010

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson, A. T., … Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17*(1), 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

Feng, X., Park, D. S., Walker, C., Peterson, A. T., Merow, C., & Papeş, M. (2019). A checklist for maximizing reproducibility of ecological niche models. *Nature Ecology and Evolution*, *3*(10), 1382–1395. https://doi.org/10.1038/s41559-019-0972-5

Guillera-Arroita, G., Lahoz-Monfort, J. J., Elith, J., Gordon, A., Kujala, H., Lentini, P. E., Mccarthy, M. A., Tingley, R., & Wintle, B. A. (2015). Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, *24*(3), 276–292. https://doi.org/10.1111/geb.12268

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, *8*(9), 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). Habitat suitability and distribution models with applications in R. *Cambridge University Press*. https://doi.org/10.1017/9781139028271

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3–4), 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., & Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, *6*(4), 366–379. https://doi.org/10.1111/2041-210X.12352

Thuiller, W., Lafourcade, B., Engler, R., & Araújo, M. B. (2009). BIOMOD - a platform for ensemble forecasting of species distributions. *Ecography*, *32*(3), 369–373. https://doi.org/10.1111/j.1600-0587.2008.05742.x

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J., & Elith, J. (2021). Predictive performance of presence-only species distribution models: A benchmark study with reproducible code. *Ecological Monographs*, *92*, e01486. https://doi.org/10.1002/ecm.1486

Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J. J., Leitão, P. J., Park, D. S., Peterson, A. T., Rapacciuolo, G., Schmatz, D. R., Schröder, B., Serra-Diaz, J. M., Thuiller, W., … Merow, C. (2020). A standard protocol for reporting species distribution models. *Ecography*, *43*(9), 1261–1277. https://doi.org/10.1111/ecog.04960

## BIOSKETCH

**Boris Leroy** is a lecturer at the Muséum National d'Histoire Naturelle of Paris. He is interested in the global change biogeography of aquatic organisms, with a particular focus on climate change and invasive alien species.

**How to cite this article:** Leroy, B. (2022). Choosing presence-only species distribution models. *Journal of Biogeography*, *00*, 1–4. https://doi.org/10.1111/jbi.14505