An Evaluation of Conventional and Emerging Methods in Presence-Only Species Distribution Modeling: A Literature Review

Benton Tripp

***Abstract* — As species distribution models (SDMs) continue to evolve and diversify, there is an increasing need for a thorough examination of both traditional and novel techniques. This review emphasizes the application of presence-only prediction methods, notably Poisson point processes and MaxEnt, within species distribution modeling. A comparative analysis is provided between traditional machine learning models (e.g., gradient boosting and random forests) and these more common presence-only prediction techniques. Additionally, this work delves into Bayesian approaches, data processing techniques, and innovative sampling methodologies, underlining various strategies that have been employed to enhance model accuracy and minimize bias. The primary aim of this review is to offer a robust foundation for future research focusing on the amalgamation and enhancement of presence-only SDM strategies.**

I. Introduction

A. Background and Significance

Species distribution modeling (SDM) has become increasingly significant in the realm of ecological research. As technology, remote sensing, and computational power have advanced, a variety of methods have emerged to predict species distributions (Renner & Warton, 2013; Marmion et al., 2008; O'Sullivan and Unwin, 2010). Species observation data is intrinsically challenging to model because it is typically "presence-only" data. That is, there are no records of observing a species' absence. Models using presence-only data confront distinct challenges due to uncertainties stemming from inherent bias and limitations within the data.

Historical modeling techniques of presence-only data largely relied on measuring environmental similarities. In contrast, contemporary methodologies place a stronger emphasis on modeling species' suitability in relation to their ambient environment. This contemporary approach utilizes what is termed "background" or "pseudo-absence" data: specific locales within the study domain where the species' presence remains undetermined but which but which represent a spectrum of environmental conditions potentially indicative of a species' presence or absence (Busby 1991, Carpenter et al. 1993; Phillips et al., 2009b).

Several presence-only modeling techniques are predominant in the literature. Poisson point processes, applied in both likelihood and Bayesian frameworks, model with solely presence points and environmental covariates, eliminating the requirement for absence or pseudo-absence points. Conversely, logistic regression and Maximum Entropy methods incorporate observation and pseudo-absence points as binary indicators, using environmental and temporal variables as predictors (Warton & Shepherd, 2010; Chakraborty et al., 2011; Ward et al., 2009; Dorazio, 2012).

Machine learning models like Random Forests (Zhang et al., 2019), Support Vector Machines, and Gradient Boosting Machines have been adapted to model presence-only data in more recent developments, employing pseudo-absence points to ascertain species' distributions. These models identify intricate data relationships, thereby refining species occurrence predictions. With ensemble or kernel-based methodologies, they adeptly capture the intricate patterns typical of ecological data (Valavi et al. 2021; Zhang et al., 2019).

While machine learning models are increasingly prevalent, Bayesian techniques provide additional depth by quantifying the aforementioned uncertainties associated with presence-only data. Notable Bayesian techniques encompass Markov Chain Monte Carlo (MCMC) and Integrated Nested Laplace Approximation with Stochastic Partial Differential Equation (INLA-SPDE). Although MCMC is well-established, INLA-SPDE offers superior computational efficiency, with both methodologies producing models that accurately quantify uncertainties (Divino, 2013; Lezama-Ochoa et al., 2020).

B. Objective

The objective of this review is to critically evaluate and synthesize the developments in presence-only prediction techniques. The exploration aims to evaluate the performance of traditional machine learning models against more specialized techniques, diving deep into the nuances of Bayesian methods, presence-only data processing mechanisms, and innovative sampling methodologies.

II. Pseudo-Absence Data in Presence-Only Models

A. Pseudo-Absence Selection Strategies

Pseudo-absence data selection significantly influences the reliability and accuracy of presence-only models. Wisz and Guisan (2009) explored various pseudo-absence selection methods for species distribution models. They examined three techniques: (a) actual absences, (b) random pseudo-absences from the background, and (c) a two-stage method that first uses either the Ecological Niche Factor Analysis (ENFA) or BIOCLIM to identify low suitability regions and then selects pseudo-absences from those regions.

Wisz and Guisan (2009) further distinguished between two core modeling techniques: "profile techniques," relying solely on species presence data (with BIOCLIM, DOMAIN, and ENFA as examples) and "group discrimination techniques," necessitating both presence and absence data, such as logistic regression. Given the frequent lack of genuine absence data,

many species distribution studies resort to pseudo-absences. A significant risk here is that randomly chosen pseudo-absences might overlap with true species presence sites, introducing potential inaccuracies.

In adopting their two-step modeling approach, Engler et al. initially utilized the profile technique "ENFA" to evaluate habitat suitability and then incorporated pseudo-absences from low suitability areas into a logistic regression model for predicting species distribution. However, Wisz and Guisan (2009) found that this two-step methodology often misaligned with the actual data. Through a virtual species method, they illustrated that models grounded in randomly selected pseudo-absences might provide superior predictive capability and variable selection over two-step methods. A critical drawback of two-step models is the potential for overfitting. While these models may align well with training data, they might underperform when generalized to new datasets.

Furthermore, Wisz and Guisan (2009) underscored the value of parsimony in model selection. They posited that models with fewer predictors frequently yield more robust predictions. Contrary to methods like BIOCLIM and ENFA, which lack mechanisms for predictor selection and can't process quadratic predictor relationships, tools like MaxEnt can dynamically adjust complexity based on data volume, often outperforming models with rigidly complex response patterns..

Senay et al. (2013) further explored pseudo-absence selection. They underscored the importance of having both presence and absence data for generating reliable SDM predictions and proposed a method that integrates both geographical and environmental factors. This three-step method first defines a boundary around species presence points based on environmental variables. Next, it identifies areas environmentally different from presence points but still within the boundary. Lastly, it uses K-means clustering to refine the selection of pseudo-absences, with cluster centroids being prioritized.

Building on the methodological innovations like those proposed by Senay et al., it's essential to understand the foundational principles and concerns surrounding SDMs. At their core, SDMs estimate species' distributions from known presence locations. While mechanistic models provide another perspective, they require in-depth knowledge of a species' environmental requirements, which isn't always available. Concerns regarding the accuracy of SDMs, particularly when absence data is unavailable or unreliable, have been voiced by various researchers (Jime´nez-Valverde et al, 2008; Phillips et al., 2006; and others).

B. Presence/Absence Data vs. Presence-Only Data

The nuances between presence/absence and presence-only data are integral to understanding the modeling process. Gelfand (2018) discussed these differences, noting that locations are fixed in presence/absence data, but random in presence-only data. This distinction leads to differences in defining the "probability of presence". Despite suggestions from earlier studies that presence/absence models can be derived from presence-only data and that the two data types can be integrated, Gelfand presents a contrasting view. He emphasizes that presence/absence data should be modeled at individual locations, using two surfaces. One indicates the probability of presence, while the other provides a binary map based on Bernoulli trials. On the other hand, presence-only data should be based on a point pattern, arising from a random count of sightings, and governed by an intensity function, unrelated to Bernoulli trials.

Gelfand (2018) also stressed the importance of consistency in modeling. The presence-only approach, using point patterns, results in intensity surfaces. These do not provide a probability of presence but indicate relative likelihoods of species sightings. He highlighted that there's a distinct difference between the likelihood of sighting a species once or multiple times. This differentiation is key to understanding species distributions in specific regions. In essence, a presence can't be reduced to a single point, but can be identified by observations within an area. As a result, presence/absence data, when mapped, presents as distinct regions or "patches" where a species resides. These patches, when observed from an ecological standpoint, represent areas densely populated by a species, with absences marked by significant gaps.

Gelfand (2018) calls for a more meticulous approach to modeling when studying species distributions. Recognizing and accounting for the inherent incompatibilities between data types is pivotal to ensure accurate inference.

III. MaxEnt and Poisson Point Process in Species Distribution Modeling

A. Theoretical Underpinnings

Maximum Entropy (MaxEnt) and Poisson Point Processes (PPMs) have been identified as essential tools for species distribution modeling. MaxEnt models the probability per grid cell and analyzes data after aggregating them into presence/absence grid cells. In contrast, a Poisson PPM models the limiting expected count or intensity per unit area, rather than per grid cell. This per area basis, compared to per grid cell, is a significant distinction between the two approaches (Renner & Warton, 2013).

Interestingly, Renner & Warton (2013) highlight the mathematical equivalence of the MaxEnt procedure and Poisson regression, asserting that both approaches fit the same model and estimate parameters to maximize the same function up to a constant. Moreover, the MaxEnt and PPM solutions for grid cell data are proportional, with identical estimates of slope parameters.

B. Applications and Limitations

MaxEnt's applications are sometimes hindered by its shortcomings. It lacks clarity regarding diagnostic tools to assess model fit and is unclear about the spatial resolution when constructing grid cells (Renner & Warton, 2013).

A key limitation of MaxEnt is its scale dependence of predicted probabilities and arbitrary choice of spatial resolution. The per grid cell analysis is not invariant under choice of spatial resolution, unlike PPM, which models intensity on a per area basis (Renner & Warton, 2013). MaxEnt also fails to estimate the intercept consistently, diverging to $-\infty$ as spatial resolution increases (Renner & Warton, 2013; Elith et al., 2011).

PPM, on the other hand, offers various solutions to MaxEnt's problems. Predicted intensities in PPM are scale-invariant, and spatial resolution can be increased until log-

likelihood converges. Various goodness-of-fit procedures are available for PPM, enabling more robust model adequacy assessment (Renner & Warton, 2013; Cressie, 1993; Baddeley et al., 2005).

Warton & Shepherd (2010) introduce PPMs as an alternative to pseudo-absence approaches, which have weaknesses in model specification, interpretation, and implementation. Point process modeling directly addresses these concerns, proposing a more sound specification for observed data without needing to generate new data. PPM also provides a framework for the selection of pseudo-absences, an area often addressed ad hoc in ecology (Warton & Shepherd, 2010).

C. Insights into Presence-Only Prediction

Presence-only prediction is a critical aspect of species distribution modeling. MaxEnt is limited in this regard due to its scale dependence and the current ambiguity over the spatial resolution (Renner & Warton, 2013).

PPM is proposed as a solution to the "pseudo-absence problem" in presence-only data, providing a more robust specification, clearer interpretation, and structured implementation than MaxEnt. The problems related to the pseudo-absence approach are rectified through the application of a point process modeling framework (Warton & Shepherd, 2010). As discussed in section II., various authors have addressed the confusion over how pseudo-absences should be chosen (Elith and Leathwick, 2007; Guisan et al., 2007; Zarnetske, Edwards, and Moisen, 2007; Phillips et al., 2009), recognizing that the selection method can yield different outcomes (Chefaoui and Lobo, 2008).

Studies have demonstrated that logistic regression slope parameters and their corresponding standard errors converge to those of the Poisson point process model as the number of pseudo-absences is increased (Warton & Shepherd, 2010). This demonstrates that the PPM approach successfully addresses the arbitrary nature of pseudo-absence selection, signifying that a specific form of point process model is being estimated, even in the utilization of pseudo-absence methods. Current selection procedures for pseudo-absences do not align with best practices, as they are frequently chosen at random and lack a basis in convergence criteria (Pearce and Boyce, 2006; Zarnetske, Edwards, and Moisen, 2007).

The exploration of MaxEnt and PPM in species distribution modeling reveals intriguing similarities and critical differences between these two methods. While MaxEnt is challenged by its shortcomings, including scale dependence and lack of consistent intercept estimation, PPM offers robust solutions, especially for presence-only prediction. The equivalence between MaxEnt and PPM and the insights into the pseudo-absence problem signify a notable contribution to ecological modeling, pointing to the potential for further refinements and innovations in the field.

IV. Machine Learning for Species Distribution Modeling

Machine learning (ML) has become increasingly relevant in species distribution modeling since the study by Elith et al. (2006) introduced innovative methods using a global dataset for SDM evaluation. Building on this work, Valavi et al. (2022) reanalyzed this dataset, comparing traditional models such as MaxEnt and generalized additive models with newer methods like XGBoost, random forests, support vector machines, and the ensemble modeling framework 'biomod'.

Their analysis revealed that ensemble models built from individually fine-tuned models performed better than those constructed using the default biomod framework. Furthermore, nonparametric ML methods such as random forests proved effective in handling complex data patterns without overfitting (Valavi et al., 2022). While there are situations where traditional regression methods are preferred, especially in cases with few occurrences, they often performed less optimally than the nonparametric models.

Zhang et al. (2019) provided a detailed exploration of the random forest algorithm in the context of species distribution modeling. Random forests use an ensemble learning approach, employing both classification tree and regression tree algorithms to create categorical and numerical species distribution maps. These algorithms utilize bootstrap samples, allowing random forests to make predictions on new data. In the context of classification trees, model outputs can serve dual purposes: as categories and as indices of occurrence (Strobl et al., 2009).

A major concern in SDM performance is the selection of evaluation metrics. Zhang et al. (2019) recommended using multiple metrics, highlighting the importance of threshold-independent evaluations such as RMSE, MAE, and AUC. Their research also addressed the challenges of converting numerical predictions into binary outcomes and emphasized the impact of the chosen threshold-setting method (Zhang et al., 2019). They identified four threshold methods—MaxKappa, MaxOA, MinROCdist, and MaxTSS—as effective for converting continuous predictions with presence-only data into binary outcomes.

Zhang et al. (2019) also proposed a framework for applying random forests with presence-only data in SDM. While this framework can be broadly applied, it stresses the importance of considering both model discrimination and reliability, suggesting that adjustments may be needed based on specific species and ecological conditions.

The studies by Valavi et al. and Zhang et al. emphasize the significant impact of the algorithm selection on SDM outcomes. Valavi et al. focused on the benefits of ensemble and nonparametric models, while Zhang et al. explored the detailed strengths and applications of the random forest algorithm. Collectively, their findings indicate a trend toward leveraging ML in SDM, underscoring its potential for future ecological research.

V. Bayesian Approaches In Species Distribution Modeling

A. Integrated Nested Laplace Approximation with Stochastic Partial Differential Equation (INLA-SPDE)

The Integrated Nested Laplace Approximation with Stochastic Partial Differential Equation (INLA-SPDE) framework offers a Bayesian technique for addressing species distribution modeling. Bayesian models, such as INLA-SPDE, effectively handle complex datasets characterized by spatial and temporal autocorrelations. They stand as an alternative to frequentist methods, which provide fixed parameter estimates

(Lezama-Ochoa et al., 2020; Martínez-Minaya et al., 2018; Blangiardo & Cameletti, 2015).

The strength of the INLA-SPDE method lies in its capacity to identify both common and less frequent areas of species presence, thus enhancing the predictive power of SDMs (Lezama-Ochoa et al., 2020). This approach factors in multilevel structures with spatial random effects, which represent diverse spatial processes influencing species patterns (Lezama-Ochoa et al., 2020; Pennino et al., 2017; Redding et al., 2017).

A key feature of INLA-SPDE is its use of Delaunay triangulation instead of the typical grids in SDMs (*see section III*). This method gathers more information in regions with a higher density of observations, yielding more precise predictions (Lezama-Ochoa et al., 2020). However, INLA-SPDE has limitations, including challenges in handling categorical variables and complexities in spatial data triangulation (Lezama-Ochoa et al., 2020). While it is faster than Markov Chain Monte Carlo (MCMC) methods, it serves better as an addition or alternative, rather than a replacement (Rue et al.; Lezama-Ochoa et al., 2020).

The Bayesian foundation of INLA-SPDE enables precise quantification of uncertainties, providing credible intervals and standard deviations in conjunction with point estimates (Lezama-Ochoa et al., 2020). This precision provides a clearer of understanding the species distribution in question. But for a more holistic analysis, it's important to compare INLA-SPDE to models such as Random Forests, MaxEnt, and Boosted Regression Trees, recognizing the unique strengths and limitations of each approach (Lezama-Ochoa et al., 2020).

B. Markov Chain Monte Carlo (MCMC)

Divino (2013) presents a structured Bayesian model specifically designed to estimate parameters of a linear logistic regression suitable for presence-only data. This model aims to connect the observed stratum variable Z with covariates X, especially when a binary response Y is missing. The absence of this response introduces two levels of uncertainty: one from the censoring mechanism and another from the sampling procedure.

Divino (2013) provides equations that detail the model's functionality. The primary equation for presence-only data is:

$$\phi_{\text{pod}}(x) \approx x\beta + \log\left(\frac{n_1 u + np}{n_1 u}\right)$$

This equation adapts the traditional linear regression model for presence-only datasets integrating linear prediction $x\beta$, (where $\phi(x) = x\beta$, and $\beta = (\beta_1, \ldots \beta_k)$ is a vector of the regression parameters), with a logarithmic correction term[1]. Central to the Bayesian approach is the Markov Chain Monte Carlo (MCMC) algorithm. Divino (2013) emphasizes the role of data augmentation within the MCMC computation, ensuring consistent value derivation for $n_1 u$ crucial for adapting the regression function for presence-only data.

Divino (2013) outlines the MCMC algorithm sequence:

1. Initialization of hyperparameters and latent variables.
2. Calculation of the sum of latent variables to adjust the regression function.
3. Sampling of hyperparameters based on the observed data.
4. Estimation of linear parameters conditional on the hyperparameters.
5. A sampling step for unobserved data in the presence-only framework.

The process is reiterated to refine estimates, thereby leveraging the strengths of Bayesian methods. Divino (2013) also summarizes the hierarchical layout for the Bayesian model as follows:

- The hyperparameter θ, at the top, governs the distribution of β.
- At the next level, the linear parameters β relates the covariates X and the response Y.
- The unobserved data y_u are modeled as latent parameters in a Bernoulli distribution.
- At the base is the likelihood related to the observable variable Z.

Through this structure, Divino (2013) manages the multiple sources of uncertainty, offering a streamlined approach for presence-only data. The combination of Bayesian modeling with the MCMC framework, as described by Divino (2013), provides a robust method for analyzing presence-only data. As this type of data becomes more relevant in fields like ecology, Divino's model serves as a foundational reference for further studies aiming for reliable predictions and analyses.

VI. DISCUSSION

A. Synthesis of Findings

The exploration of presence-only species distribution modeling reveals many subtle challenges present in ecological research. Traditional models based on environmental similarities have given way to more complex methodologies that highlight species suitability in correlation with their ambient environment (Wisz & Guisan, 2009). Notably, the shift from logistic regression and Maximum Entropy to Poisson Point Process models signifies the endeavor to overcome the challenges posed by presence-only data, particularly in the realm of pseudo-absence data (Renner & Warton, 2013). Machine learning techniques, especially ensemble models, also offer greater precision and adaptability in species distribution modeling (Valavi et al., 2022; Zhang et al., 2019). Bayesian techniques such as INLA-SPDE and Markov Chain Monte Carlo (MCMC) have also emerged as promising tools capable of addressing uncertainties inherent in presence-only data (Divino, 2013).

B. Implications for Future Research

There are still unresolved challenges warranting focused research in the domain of presence-only prediction. The

[1] The proposition, framework, and proof underlying the logarithmic correction term are detailed by Divino (2013). However, an in-depth exploration of these aspects is beyond the scope of this literature review.

discrepancies between presence/absence and presence-only data, as elucidated by Gelfand (2018), necessitate further research into innovative sampling methodologies and coherent modeling techniques. While MaxEnt remains relevant, its challenges highlight the exigency of refining existing models and the potential of PPM as a robust alternative (Renner & Warton, 2013). Additionally, as machine learning solidifies its place in SDM, future efforts could probe deeper into fine-tuning ensemble models for superior accuracy. Lastly, while Bayesian methods such as INLA-SPDE and MCMC offer novel solutions, more comprehensive studies are required to fully harness their capabilities and address challenges like managing categorical variables (Divino, 2013).

VII. Conclusion

A. Summary of Major Findings

Presence-only species distribution modeling is marked by a significant transition from traditional models to modern, multifaceted methodologies. Pseudo-absence data, pivotal to the accuracy of SDM, poses challenges which techniques like Poisson Point Process models aim to address (Wisz & Guisan, 2009; Renner & Warton, 2013). Machine learning models, especially ensemble models, have been identified as suitable alternative methods to "traditional" species distribution modeling techniques (Elith et al., 2006; Valavi et al., 2022). Bayesian techniques have also emerged as a powerful alternative, addressing the complexities of spatial data and the challenges of addressing uncertainty in modeling presence-only data (Divino, 2013).

B. Recommendations

Given the prevailing challenges and advancements in presence-only prediction, it is imperative for researchers to:

1. Further refine and innovate sampling methodologies, ensuring consistency in modeling, especially in distinguishing between presence/absence and presence-only data (Gelfand, 2018).
2. Diversify applications and comparative studies of models like PPM and MaxEnt, weighing their strengths against their inherent challenges.
3. Augment research into the potential of machine learning, focusing on the adaptability and precision of ensemble models in SDM.
4. Expand upon the capabilities of Bayesian methods, aiming to devise more efficient and holistic approaches.

References

Aarts, G., J. Fieberg, and J. Matthiopoulos (2012). Comparative interpretation of count, presence/absence and point methods for species distribution models. Methods in Ecology and Evolution 3, 177–187.

Baddeley, A., & Turner, R. (2005). spatstat: An R Package for Analyzing Spatial Point Patterns. Journal of Statistical Software, 12(6). https://doi.org/10.18637/jss.v012.i06

Brieuc, M. S. O., Waters, C. D., Drinana, D. P., & Naish, K. A. (2018). A practical introduction to random forest for genetic association studies in ecology and evolution. Mol. Ecol. Resour., 18, 755–766.

Chefaoui, R. M., & Lobo, J. M. (2008). Assessing the effects of pseudo-absences on predictive distribution model performance. Ecological Modelling, 210(4), 478–486. https://doi.org/10.1016/j.ecolmodel.2007.08.010

Cressie, N. (1993). Statistics for spatial data. In John Wiley & Sons, Inc. eBooks. https://doi.org/10.1002/9781119115151

Divino, F. (2013, May 6). Bayesian modeling and MCMC computation in linear logistic regression for presence-only data. arXiv.org. https://arxiv.org/abs/1305.1232

Dorazio, R. M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. Global Ecology and Biogeography 23, 1472–1484.

Elith J., Graham C. H., Anderson R. P., Dudik M., Ferrier S., et al. (2006) Novel methods improve prediction of species; distributions from occurance data. Ecography 29: 129–151.

Elith, J., & Leathwick, J. R. (2009). Species Distribution Models: Ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics, 40(1), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2010). 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

Engler, R., Guisan, A., & Rechsteiner, L. (2004). An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal of Applied Ecology, 41(2), 263–274. https://doi.org/10.1111/j.0021-8901.2004.00881.x

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv., 24, 38–49.

Gelfand, A. E. (2018, September 5). Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. arXiv.org. https://arxiv.org/abs/1809.01322

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P., Tulloch, A. I. T., Regan, T. J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T. G., Rhodes, J. R., Maggini, R., Setterfield, S. A., Elith, J., Schwartz, M. W., Wintle, B. A., Broennimann, O., Austin, M. P., . . . Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. Ecology Letters, 16(12), 1424–1435. https://doi.org/10.1111/ele.12189

Hastie, T. and W. Fithian (2013). Inference from presence-only data; the ongoing controversy. Ecography 36, 864–867.

Hengl T. (2009) A Practical Guide to Geostatistical Mapping. Open Access Publication. Available: http://spatial-analyst.net/book/system/files/Hengl_ 2009_GEOSTATe2c1w.pdf.

Hirzel A. H., Hausser J., Chessel D, Perrin N (2002) Ecological-Niche factor analysis: How to compute habitat-suitability maps without absence data? Ecology 83: 2027.

Jime´nez-Valverde A., Lobo J. M., Hortal J. (2008) Not as good as they seem: the importance of concepts in species distribution modelling. Diversity and Distributions 14: 885–890.

Kearney M., Porter W. (2009) Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. Ecology Letters 12: 334.

Lezama-Ochoa, N., Pennino, M. G., Hall, M., Lopez, J., & Murua, H. (2020). Using a Bayesian modelling approach (INLA-SPDE) to predict the occurrence of the Spinetail Devil Ray (Mobular

mobular). Scientific Reports, 10(1). https://doi.org/10.1038/s41598-020-73879-3

Li W., Guo Q., Elkan C. (2011) Can we model the probability of presence of species without absence data? Ecography 34: 1096–1105.

Liu, C., White, M., & Newell, G. (2011). Measuring and comparing the accuracy of species distribution models with presence–absence data. Ecography, 34, 232–243.

Lobo J. M., Jime´nez-Valverde A., Hortal J. (2010) The uncertain nature of absences and their importance in species distribution modelling. Ecography 33: 103–114.

Lorena A. C., Jacintho L. F. O., Siqueira M. F., Giovanni R. D., Lohmann L. G., et al. (2011) Comparing machine learning classifiers in potential distribution modelling. Expert Systems with Applications 38: 5268–5275.

Marmion, M., Hjort, J., Thuiller, W., & Luoto, M. (2008). A comparison of predictive methods in modelling the distribution of periglacial landforms in Finnish Lapland. Earth Surface Processes and Landforms, 33(14), 2241–2254. https://doi.org/10.1002/esp.1695

Martínez-Minaya, J., Cameletti, M., Conesa, D. & Pennino, M. G. Species distribution modeling: a statistical review with focus in spatio-temporal issues. in Stochastic Environmental Research and Risk Assessment 1–18 (2018).

O'Sullivan, D., & Unwin, D. (2010). Geographic Information Analysis. https://doi.org/10.1002/9780470549094

Pacifici, K., B. J. Reich, D. A.W. Miller, B. Gardner, G. Stauffer, S. Singh, A. McKerrow, and J. A. Collazo (2017). Integrating multiple data sources in species distribution modeling: a framework for data fusion. Ecology, 840–850.

Pearce, J., & Boyce, M. S. (2006). Modelling distribution and abundance with presence-only data. Journal of Applied Ecology, 43(3), 405–412. https://doi.org/10.1111/j.1365-2664.2005.01112.x

Pearce, J., & Ferrier, S. (2000). Evaluating the predictive performance of habitat models developed using logistic regression. Ecol. Modell., 133, 225–245.

Pennino, M. G., Vilela, R., Bellido, J. M. & Mendoza, M. Comparing methodological approaches to model occurrence patterns of marine species. in Research Advances in Marine Resources (Eds: Norton, K.). (Nova Publisher, ISBN: 978-1-53612-177-3, 2017).

Peters, J., Baets, B. D., Verhoest, N. E. C., Samson, R., Degroeve, S., Becker, P. D., & Huybrechts, W. (2007). Random forests as a tool for ecohydrological distribution modelling. Ecol. Model., 207, 304–318.

Peterson A. T. (2006) Uses and Requirements of Ecological Niche Models and Related Distributional Models. Biodiversity Informatics 3: 59–72.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J. R., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications, 19(1), 181–197. https://doi.org/10.1890/07-2153.1

Phillips S. J., Anderson R. P., Schapire RE (2006) Maximum entropy modeling of species geographic distributions. Ecological Modelling 190: 231–259.

Redding, D. W., Lucas, T. C., Blackburn, T. M. & Jones, K. E. Evaluating Bayesian spatial methods for modelling species distributions with clumped and restricted occurrence data. PLoS ONE 12, e0187602 (2017).

Renner, I., & Warton, D. I. (2013). Equivalence of MAXENT and Poisson Point process models for species distribution modeling in Ecology. Biometrics, 69(1), 274–281. https://doi.org/10.1111/j.1541-0420.2012.01824.x

Royle, J. A., R. B. Chandler, C. Yackulic, and J. D. Nichols (2012). Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. Methods in Ecology and Evolution 3, 545–554.

Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Ser. B. (Stat. Method.) 71, 319–392 (2009).

Senay SD, Worner SP, Ikeda T (2013) Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modelling. PLoSONE 8(8): e71218. doi:10.1371/journal.pone.0071218

Soberon J., Peterson A. T. (2005) Interpretation of models of fundamental ecological niches and species' distriputional areas. Biodiversity Informatics 2: 1–10.

Strobl, C., Malley, J. D., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application and characteristics of Classification and Regression Trees, Bagging and Random Forests. Psychol. Methods, 14, 323–348.

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(1). https://doi.org/10.1002/ecm.1486

Wisz, M. S., & Guisan, A. (2009). Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. BMC Ecology, 9(1), 8. https://doi.org/10.1186/1472-6785-9-8

Warton, D. I., & Shepherd, L. (2010). Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. The Annals of Applied Statistics, 4(3). https://doi.org/10.1214/10-aoas331

Zarnetske, P. L., Edwards, T. C., & Moisen, G. G. (2007). HABITAT CLASSIFICATION MODELING WITH INCOMPLETE DATA: PUSHING THE HABITAT ENVELOPE. Ecological Applications, 17(6), 1714–1726. https://doi.org/10.1890/06-1312.1

Zhang, L., Huettmann, F., Zhang, X., Liu, S., Park, S., Yu, Z., & Mi, C. (2019). The use of classification and regression algorithms using the random forests method with presence-only data to model species' distribution. MethodsX, 6, 2281–2292. https://doi.org/10.1016/j.mex.2019.09.035

Zhang, L., Wang, L., Liu, S., Sun, P., Yu, Z., Huang, S., & Zhang, X. (2017). An evaluation of four threshold selection methods in species occurrence modelling with random forest: case studies with Davidia involucrata and Cunninghamia lanceolata. Chin. J. Plant Ecol., 41, 387–395.