ECOGRAPHY

Review

Optimising occurrence data in species distribution models: sample size, positional uncertainty, and sampling bias matter

Vítězslav Moudrý[1], Manuele Bazzichetto[1], Ruben Remelgado[2,3], Rodolphe Devillers[4], Jonathan Lenoir[5], Rubén G. Mateo[6], Jonas J. Lembrechts[7], Neftalí Sillero[8], Vincent Lecours[9], Anna F. Cord[2,3], Vojtěch Barták[1], Petr Balej[1], Duccio Rocchini[1,10], Michele Torresani[11], Salvador Arenas-Castro[12], Matěj Man[13], Dominika Prajzlerová[1], Kateřina Gdulová[1], Jiří Prošek[1,13], Elisa Marchetto[10], Alejandra Zarzo-Arias[14,15], Lukáš Gábor[1], François Leroy[1], Matilde Martini[10], Marco Malavasi[16], Roberto Cazzolla Gatti[10], Jan Wild[1,13] and Petra Šímová[1]

[1]Department of Spatial Sciences, Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Praha-Suchdol, Czech Republic
[2]Chair of Computational Landscape Ecology, TUD Dresden University of Technology, Dresden, Germany
[3]Agro-Ecological Modeling Group, Institute of Crop Science and Resource Conservation, University of Bonn, Bonn, Germany
[4]UMR Espace-Dev, Institut de Recherche Pour le Développement, Univ Réunion, La Réunion, France
[5]UMR CNRS 7058 'Ecologie et Dynamique des Systèmes Anthropisés' (EDYSAN), Université de Picardie Jules Verne, Amiens, France
[6]Departamento de Biología and Centro de Investigacion en Biodiversidad y Cambio Global (CIBC-UAM), Universidad Autonoma de Madrid, Madrid, Spain
[7]Research Group of Plants and Ecosystems (PLECO), Department of Biology, University of Antwerp, Antwerp, Belgium
[8]Centro de Investigação em Ciências Geo-Espaciais (CICGE), Faculdade de Ciências da Universidade do Porto, Alameda do Monte da Virgem, Vila Nova de Gaia, Portugal
[9]Université du Québec à Chicoutimi, Saguenay, QC, Canada
[10]BIOME Lab, Department of Biological, Geological and Environmental Sciences, Alma Mater Studiorum University of Bologna, Bologna, Italy
[11]Free University of Bolzano/Bozen, Faculty of Agricultural, Environmental and Food Sciences, Bolzano/Bozen, Italy
[12]Área de Ecología, Dpto. de Botánica, Ecología y Fisiología Vegetal, Facultad de Ciencias, Universidad de Córdoba, Edificio Celestino Mutis (C-4), Córdoba, Spain
[13]Institute of Botany of the Czech Academy of Sciences, Průhonice, Czech Republic
[14]Universidad de Oviedo, Oviedo, Spain
[15]Department of Biogeography and Global Change, Museo Nacional de Ciencias Naturales (MNCN-CSIC), Madrid, Spain
[16]Department of Chemistry, Physics, Mathematics and Natural Sciences, University of Sassari, Sassari, Italy

Correspondence: Manuele Bazzichetto (manuele.bazzichetto@gmail.com)

Ecography
2024: e07294
doi: 10.1111/ecog.07294
Subject Editor: Miguel Araújo
Editor-in-Chief: Miguel Araújo
Accepted 1 July 2024

NORDIC SOCIETY OIKOS
www.ecography.org

Species distribution models (SDMs) have proven valuable in filling gaps in our knowledge of species occurrences. However, despite their broad applicability, SDMs exhibit critical shortcomings due to limitations in species occurrence data. These limitations include, in particular, issues related to sample size, positional uncertainty, and sampling bias. In addition, it is widely recognised that the quality of SDMs as well as the approaches used to mitigate the impact of the aforementioned data limitations depend on species ecology. While numerous studies have evaluated the effects of these data limitations on SDM performance, a synthesis of their results is lacking. However, without a comprehensive understanding of their individual and combined effects, our ability to predict the influence of these issues on the quality of modelled species–environment associations remains largely uncertain, limiting the value of model outputs. In this paper, we review studies that have evaluated the effects of sample size, positional uncertainty, sampling bias, and species ecology on SDMs outputs. We build upon their findings to provide recommendations for the critical assessment of species data intended for use in SDMs.

Keywords: data quality, ecological niche modelling, filtering, sampling, spatial scale, validation

Introduction

The quantity and quality of biological observations have improved dramatically over the past few decades. However, a certain level of uncertainty is inherently present in such data, resulting in uncertainties of scientific inferences based on them (Hortal et al. 2015, Daru and Rodriguez 2023, Hughes et al. 2023). Correlative species distribution models (SDMs; also known as habitat suitability models or ecological niche models; Sillero 2011) are useful for tackling the gaps in our knowledge of species occurrence (Elith and Leathwick 2009). These models combine environmental and species occurrence data to build a set of rules describing the environmental space where species were observed (i.e. species ecological niche) and can then be used to predict the distribution of that species (Ferrier et al. 2017). SDMs support a wide variety of ecological applications, such as the assessment of the spread of invasive species (Guisan et al. 2013, Bazzichetto et al. 2021), the detection of potential impacts of environmental changes on biodiversity (Ehrlén and Morris 2015, Haesen et al. 2023), or the identification of suitable locations for the relocation of endangered species (Guisan et al. 2013, Segal et al. 2021). However, despite their broad applicability, SDMs have critical shortcomings associated in particular with the characteristics of input data, including their quantity and quality (Elith et al. 2002, Barry and Elith 2006, Rocchini et al. 2011, Moudrý and Šímová 2012, Wüest et al. 2020, Davies et al. 2023). In this paper, we focus on the limitations of species occurrence data (for issues associated with environmental data, see for example Fourcade et al. 2018, Araújo et al. 2019 Moudrý et al. 2023).

Limitations of species occurrence data can introduce uncertainty and biases in the estimation of species–environment relationships and, consequently, of their predicted distributions (Araújo et al. 2019). In particular, data availability (i.e. sample size) is critical; the smaller the minimum sample size that can theoretically be used in SDMs, the higher the number of species that can be modelled (Stockwell and Peterson 2002). However, measurement errors associated with data acquisition methods (i.e. positional error; Smith et al. 2023) are another major source of uncertainty, which may, in effect, necessitate the use of a larger sample size than had the data been accurate. In addition, the choice of inappropriate sampling strategies can introduce biases towards certain locations (i.e. sampling bias; Bazzichetto et al. 2023). Moreover, it is well-recognised that the quality of SDMs is also influenced by the species' ecology (Segurado and Araujo 2004, Heikkinen et al. 2006, Guisan et al. 2007, McPherson and Jetz 2007, Collart et al. 2023) and the fact that the effects of different data limitations (e.g. sample size, positional uncertainty, and sampling bias) may be species-specific.

As the interest in using SDMs continues to grow, tackling data limitations becomes increasingly critical (Araújo et al. 2019, Wüest et al. 2020, Jansen et al. 2022, Marcer et al. 2022). In this context, it is now expected that data characteristics and limitations are considered and properly reported during the conceptualisation and validation of SDMs (Feng et al. 2019, Zurell et al. 2020, Sillero and Barbosa 2021, Tessarolo et al. 2021, Jansen et al. 2022, Jeliazkov et al. 2022, Boyd et al. 2023). However, without proper knowledge of the individual or combined effects of sample size, positional uncertainty, sampling bias, and their interaction with species' ecology, our ability to anticipate the impact of these issues on the quality of SDMs remains largely uncertain, limiting the value of model outputs (see Fig. 1 for a diagram introducing data characteristics and their relationships considered in this review).

A common approach to the evaluation of the effects of data limitations on model performances is to manipulate the input data experimentally or to simulate datasets impacted by various sources of bias or uncertainty. Here, we examine studies that manipulated sample size (section 'Sample size') or introduced positional uncertainty (section 'Positional uncertainty') or sampling bias (section 'Sampling bias') to investigate their impact on SDMs' outputs. Building upon these studies, we provide guidance on how to critically assess the spatial data used in SDMs, and identify directions for optimising the tradeoffs between data limitations and accurate modelling of species–environment relationships (section 'Guidelines and future directions').

Sample size

Among all possible factors, sample size (Box 1) has the most profound effect on the performance of an SDM (Thibaud et al. 2014, Santini et al. 2021). Sample size poses an important constraint to the model complexity, i.e. to the number of parameters to be estimated, as well as to the algorithms and their settings used for modelling. In SDMs, sample size can range from just a few (Papeş and Gaubert 2007, Pearson et al. 2007) to millions (Botella et al. 2023, Gábor et al. 2024) of records. In the vast literature measuring the effect of sample size on model performance (Table 1), the primary concern has been to determine the minimum adequate sample size required to produce reliable and fit-for-purpose models (Stockwell and Peterson 2002, Hanberry et al. 2012, Proosdij et al. 2016). In parallel, ecological research investigates to what extent additional time and economic

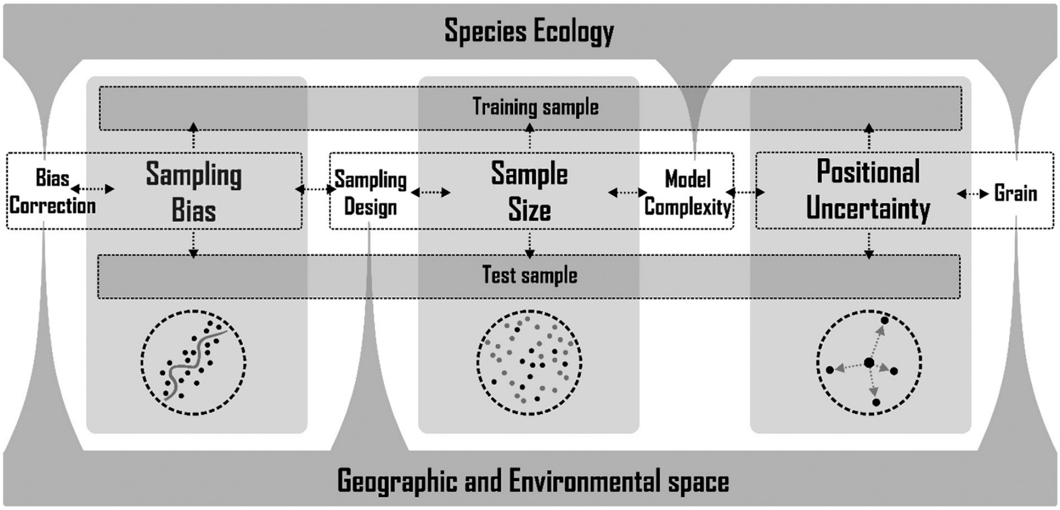

Figure 1. Sample size, positional uncertainty, and sampling bias are the three essential characteristics of species occurrence data addressed in this review. These interconnected characteristics can have a significant impact on the reliability of species distribution models (SDMs) results. Researchers must thoughtfully address these factors during the collection of species occurrence data (sampling design) and the formulation of models (model complexity). Maximising sample size, using sampling bias correction methods, and minimising positional uncertainty relative to the spatial resolution and autocorrelation of environmental predictors during model training and testing, are all essential steps. Additionally, species ecology and the distribution of species observations in the geographic and environmental space can exacerbate or attenuate the negative effects of small sample size, high sampling bias, and high positional uncertainty on the reliability of SDMs results. See Box 1 for definitions of key terms and concepts.

resources should be spent to improve models by increasing the sample size (Liu et al. 2019). Knowing the minimum (and maximum) sample size required for accurate predictions would theoretically allow optimisation of the resources spent on labour-intensive fieldwork and, therefore, help reduce associated costs. Nonetheless, the extent to which modelling could replace fieldwork remains questionable.

Importance of sample size in model training and testing

Studies focusing on a better understanding of how sample size impacts models' accuracy revealed that it is in principle possible to train SDMs with a relatively small sample. Values typically range from 50 to 150 presences (or presences–absences), although values as low as 10 presences or as high as a few hundred have also been reported (Table 1). However, it is important to note that studies typically reported minimum sample size when the model was still relatively useful, not sample size when the model gave optimal results. Besides, it has been reported that models relying on fewer than approximately 70 presences do not reliably identify the variables affecting distributional patterns (Smith and Santos 2020) or result in poor(er) estimates of the shapes of species response curves (Coudun and Gégout 2006, Shiroyama et al. 2020, Bazzichetto et al. 2023, Wang and Jackson 2023). In general, all studies agreed that increasing sample size increased a model's predictive performance (keeping the number of predictors fixed), although a plateau in model performance is generally reached (Stockwell and Peterson 2002). According to recent studies, hundreds of presences are needed to reach the plateau where increasing sample size further adds little to the model performance (Liu et al. 2019, Gábor et al. 2020a).

Insufficient attention has so far been devoted to the evaluation of possible effects of the testing dataset sample size on validating SDMs' predictive performances. Generally, small validation datasets can lead to inaccurate assessment of model performance (Hallman and Robinson 2020). Recently, Jiménez-Valverde (2020) showed that 30 presence–absence records (i.e. 15 presences and 15 absences) are a (minimum) adequate sample size for a validation dataset to estimate the predictive performance of presence–absence models. However, their conclusions are based on simulations, and it is important to note that studies using real data are essential to generalise these results. In addition, the minimal sample size of a validation dataset has not yet been evaluated in the case of presence–background data; since these carry less information than presence–absence data, the validation set should be correspondingly larger (Collart and Guisan 2023). While the importance of a sufficiently large validation sample is intuitive, the impact of increasing the sample size of the testing dataset on validation accuracy urgently needs further testing.

Relationships between sample size, species ecology, and model complexity

The association between model performance and sample size depends largely on the species' ecology. Studies have repeatedly indicated that, for a given sample size, SDMs better predict species with restricted geographical distributions (i.e. low range size, prevalence, or relative occurrence

Box 1. Glossary of key terms.

Ecological niche: Hutchinsonian niche, defined as a hypothetical hypervolume spanned by the eco-physiological responses of a species to all environmental factors affecting its fitness.

Model complexity refers to the level of intricacy and flexibility in the representation of a species' ecological niche. It reflects how well the model can capture the underlying relationships between predictors and species distribution. The choice of model complexity depends on the nature of the problem, the amount and quality of available data, the number of model parameters, and the available computational resources. Finding the right balance between a model's ability to capture patterns and its potential for overfitting is a key challenge in building effective models.

Model performance: here intended in a broad sense as a model capacity of recovering the underlying species–environment relationship using available data ('explanatory' performance), while also being able to extend (predict) out of the sample used for training/calibration ('predictive' performance).

Model training is the process of teaching a machine learning or statistical model to make predictions based on data. It is a crucial step in building and developing predictive models. Model training involves using a dataset with known outcomes to enable the model to learn the underlying patterns and relationships in the data.

Model testing, also known as model evaluation, is the process of assessing the performance and effectiveness of a machine learning or statistical model using a separate (independent) dataset that the model has not seen during training. The primary purpose of model testing is to determine how well the trained model generalises to new, unseen data and to assess its predictive accuracy and reliability.

Positional uncertainty (sometimes also referred to as positional error) in species occurrence data refers to inaccuracies or uncertainty in the recorded coordinates of where a species was observed or collected. This error can result from factors such as imprecise global navigation satellite systems (GNSS) measurements, data entry mistakes, or a lack of accurate location information.

Spatial resolution or grain refers to the level of detail or granularity at which data are collected, represented, or analysed in a spatial context. It can also be thought of as the size of the smallest spatial unit in a dataset (i.e. pixel size).

Sampling design refers to the approach used to collect species occurrence data. The sampling design is a crucial aspect of SDMs, as it should in principle ensure that the data include all relevant information to represent the ecological niche of the species and the environmental conditions in the study area. The quality and representativeness of the data collected directly impact the accuracy and reliability of the model.

Sample size: the size of the data sample used to train and validate the model. Here, we define sample size as the total number of presences and absences (i.e. presence–absence data). When discussing studies based on presence–background data, we refer specifically to the number of presences.

Sampling bias: species occurrence records typically exhibit spatial bias, wherein some locations or environmental conditions are more intensively sampled than others. People sample accessible locations more intensively than remote or unpopular ones. This type of bias means that the available data used as the response variable fail to represent the complete niche of the species.

area), as well as specialist species with strict ecological requirements (i.e. narrow ecological niche) than species with wide geographic ranges and generalist (i.e. wide ecological niche) species (Stockwell and Peterson 2002, Seoane et al. 2005, Hernandez et al. 2006, Tsoar et al. 2007, Mateo et al. 2010, Tessarolo et al. 2014, Proosdij et al. 2016, Hallman and Robinson 2020, Arenas-Castro et al. 2022, Wang and Jackson 2023). The association between model performance, sample size, and species ecology can be explained by niche completeness (i.e. the proportion of a species' niche covered by the sampling). For example, if a species has a restricted ecological niche (or range), the niche may likely be well represented by a low number of occurrences. On the other hand, a large sample size does not necessarily mean a complete coverage of the entire ecological niche for widespread species (Bazzichetto et al. 2023, Boyd et al. 2023).

This is further related to model complexity. Selecting a model with an appropriate level of complexity, which would prevent overfitting noise in the data and, at the same time, allow discrimination of influential predictors from uninfluential ones and accurately capture the true species–environment relationship, remains a challenge (Merow et al. 2014, García-Callejas and Araújo 2016, Baartman et al. 2020). Building models with complex species response shapes and/or too many predictors can result in difficulties in recognising true complexity from noise, especially in case of low sample size. However, even large sample sizes can result in low accuracy in the estimation of model parameters if the model is overly complex (i.e. includes too many parameters or interactions, e.g. Wisz et al. 2008, Moreno-Amat et al. 2015). At the same time, underfitting models that are not flexible enough to describe species–environment associations risk failing to identify the factors shaping species distributions. While adding more predictor variables avoids neglecting important ones and can improve model performance, the ability to distinguish between influential and uninfluential variables depends on sample size (Smith and Santos 2020). It is, therefore, recommended to keep

Table 1. Overview of studies testing the role of the number of presences, or presences and absences, for model performance. PA, presences–absences.

Study	Number of species	Training sample	Testing sample	Study extent / resolution	No. predictors	No. obs. suggested
Stockwell and Peterson 2002	130 birds	1–100	1000; presence–background	Mexico / 3 × 3 minutes	8	At least 50 presences
Kadmon et al. 2003	192 plants	10–200	96 plots; presence–absence	Israel / 1 × 1 km	3	50–75 presences
Hernandez et al. 2006	18 animals	5–100	50 presences	California / 1 × 1 km	10	50–75 presences
Wisz et al. 2008	46 plants, animals	10–100	Presence–absence data	Five regions / 100 × 100 m; 1 × 1 km	11–13	At least 30 presences
Mateo et al. 2010	2 plants	9–60	Compared to maps created with full datasets	Ecuador / 1 × 1 km	19	At least 20 presences
Feeley and Silman 2011	65 plants	25–150	Compared to maps created with full datasets	Tropical South America / 5 × 5 km	3	Larger than evaluated
Hanberry et al. 2012	16 trees	30–2500	Presence samples not used for training	46 000 km^2 / 310 000 polygons	16	At least 200 presences
Proosdij et al. 2016	6 virtual	3–50	Compared with actual virtual species distribution	18 000 000 km^2 / 5 × 5 minutes	15	14–25 presences
Liu et al. 2019	1800 virtual	20–640	3000 presences and absences of virtual species distribution	62 500 km^2 / 1 × 1 km	6	A few hundred presences
Støa et al. 2019	30 insects	5–320	Compared to maps created with full datasets	Norway / 1 × 1 km	2	10–15 presences
Smith and Santos 2020	1 virtual	8–1024	400 presences and absences of virtual species distribution	Virtual landscape / 1024 × 1024 cells	1	At least 128 presences
McPherson et al. 2004	7 birds	50–500	500 presences–absences	South Africa / 0.25 × 0.25 degrees	61	300 PA
Coudun and Gégout 2006	54 virtual	50–5000	Not used	Not relevant	1	At least 50 PA
Jiménez-Valverde et al. 2009	1 virtual	182–182, 288	Compared with actual virtual species distribution	6576 km^2 / 0.04 × 0.04°	4	At least 70 PA
Shiroyama et al. 2020	Bluegill	50–900	110 presences absences	Seven rivers in Kanto region, Japan.	4	At least 400 PA
Bazzichetto et al. 2023	2 virtual	200–500	Compared with actual virtual species distribution	10 794 km^2 / 1 × 1 km	2	At least 200 PA
Wang and Jackson 2023	16 virtual	50–800	50 presences–absences	140 000 km^2 / 4 × 4 km	2	At least 100 PA

the number of predictors reasonably small with respect to the sample size (Williams et al. 2012, Brun et al. 2020, Ramampiandra et al. 2023). The minimum required sample size increases with the number of parameters, which also determines the complexity of the assumed species response curves (e.g. quadratic response curves or statistical interactions among predictors; Austin 2002, Barry and Elith 2006, Maggini et al. 2006, Ficetola et al. 2014, Merow et al. 2014, Bell and Schlaepfer 2016, Carretero and Sillero, 2016). To minimise the risks of overfitting and underfitting, it is useful to evaluate models with varying levels of complexity and sample size and to select the one with the best performance while also minimising the performance difference between model training and testing (Merow et al. 2014, Ramampiandra et al. 2023).

The minimum ratio of events to predictor variables is suggested by the 'events per variable' (EPV) rule. A popular criterion says that one should rely on at least ten observations per predictor considering the event class (presence or absence in case of binary data) with the lowest abundance (e.g. a dataset with 70 presences and 30 absences would allow including a maximum of three predictors; Reineking and Schröder 2006). However, it is worth noting that the EPV rule is a guideline rather than a strict rule, and it is increasingly being questioned (van Smeden et al. 2019). For example, the appropriate ratio may vary depending on the specific context and the complexity of the data (García-Callejas and Araújo 2016). Therefore, in addition to sample size, it is important to consider model complexity with respect to sampling bias and positional uncertainty (see sections 'Positional uncertainty' and 'Sampling bias').

Recommendations associated with sample size

The above-mentioned studies showed that SDMs can perform relatively well even with small sample sizes (Table 1). However, the studies mentioned in Table 1 are difficult to compare due to the use of different species, differences in the used modelling algorithms, numbers of parameters, spatial resolutions, and geographical extents. Whether the sample size is considered small or sufficient depends largely on the number of predictors in the model, and the complexity and nature of the species–environment relationships (Merow et al. 2014, Smith and Santos 2020, Bazzichetto et al. 2023). Hence, given how context-dependent these relationships are, we cannot recommend a specific threshold of what a 'small' or 'large' sample is, but we provide a series of steps that researchers should consider when preparing SDMs:

- First, the sample size required for a particular analysis requires careful consideration of the purpose of the study (Foody 2011). On the one hand, models based on low sample sizes can help identify potential knowledge gaps and optimise the allocation of funds for field surveys (e.g. to pinpoint areas with a high potential for discovering unknown populations of the studied species; Raxworthy et al. 2003, Fois et al. 2015, 2018, Rhoden et al. 2017, Becker et al. 2022). On the other hand, healthy scepticism remains in the scientific community of (macro)ecologists and biogeographers regarding the usability of predictions derived from models with small sample sizes as guidelines for applications such as modelling species ranges, predicting responses to climate change, or planning conservation efforts (Loiselle et al. 2008, Feeley and Silman 2011, Duputié et al. 2014, Muscatello et al. 2021).
- Second, species' ecology has to be considered as SDMs better predict specialist species with narrow ecological niches than generalist species with wider ecological niches (Tsoar et al. 2007).
- Third, researchers should consider the number of predictors investigated. As the ability to differentiate between influential and non-influential variables decreases with decreasing sample sizes, the challenge lies in the a priori identification of variables that genuinely influence species distribution (Smith and Santos 2020). Studies that include a small number of variables selected based on expert opinion will generally require a smaller sample size than studies that select variables from a large pool using automated algorithms (Ficetola et al. 2014).
- Fourth, the complexity of the shape of species response curves must be taken into account as models based on small sample sizes result in less precise estimates of these shapes (Bazzichetto et al. 2023). Models aiming at generating simple response curves (e.g. linear, hinge, or step) can be developed with relatively low sample sizes. However, models identifying more complex shapes such as Gaussian or even non-parametric smooth functions require much larger sample sizes. Adding interactions between variables increases the requirements for the sample size even more.
- Fifth, we cannot suggest a minimum number of presences (presences–absences) as a rule of thumb, but if a researcher is unsure whether the sample size is sufficient given the objectives and complexity of the model, we recommend testing the effect of sample size. Start with the most comprehensive model you think is appropriate in your particular case and progressively increase the sample size until you reach your possible maximum (i.e. all presences you have), and see if your model performance is reaching a plateau. If no plateau is reached, it is likely that more presences are necessary. In such a case, a reduction in the number of variables or the complexity of the response curves should be considered. Remember to set aside at least 30 presence–absence records for model validation, as recommended by Jiménez-Valverde (2020).
- Finally, while it is possible to design accurate SDMs with a well-balanced sampling of as few as 50 presences (Table 1), most observational data are too ad hoc and far from being representative of spatial variation in species–environment associations due to confounding effects of data limitations such as positional uncertainty (section 'Positional uncertainty'), or sampling bias (section 'Sampling bias'). Hence, researchers should also consider these data limitations before attempting to build a model based on a small sample.

Positional uncertainty

Species occurrence data are always prone to positional uncertainty, i.e. the difference between the actual and recorded location of a species in the coordinate reference system of the dataset. The magnitude of the positional uncertainty associated with species observations can range from a few centimetres up to tens of kilometres. Under high positional uncertainty, SDMs using environmental layers at spatial resolutions finer than the magnitude of the positional uncertainty (e.g. environmental layers at a 10 m resolution and a 50 m positional uncertainty of species observations) can estimate erroneous/misleading species–environment relationships. The potential effect of positional uncertainty on SDMs performance is determined by several interacting factors (Fig. 2). Therefore, positional uncertainty should be assessed before calibrating and validating SDMs, as it can negatively affect training and testing datasets as well as modelling decisions, such as the spatial resolution of environmental variables.

How to address positional uncertainty in training and testing datasets

Several studies have examined the impact of positional uncertainty on SDMs performance by simulating shifts in species presences (Table 2). These studies typically compare SDMs outcomes based on data with high positional accuracy against results obtained using the same data but affected by positional uncertainty of different magnitudes. Findings from these studies have been somewhat mixed: some found little effect of positional uncertainty and reported that SDMs were relatively robust to it (Graham et al. 2008, Fernandez et al. 2009); others concluded that species occurrence data with positional uncertainty generally lead to less accurate SDMs (Johnson and Gillingham 2008, Osborne and Leitao 2009, Mitchell et al. 2017).

In real-world applications, a mix of high- and low-accuracy distribution data is the most common situation, and researchers usually have to find a compromise between positional uncertainty and sample size (Smith et al. 2023). However, studies focusing on this issue yielded somewhat conflicting results. Reside et al. (2011) warned that increasing the sample size by incorporating historic species occurrence data with inaccurate positions can reduce SDMs performances. On the other hand, Smith et al. (2023) showed that the removal of data with high positional uncertainty can excessively reduce the sample size and, thus, the model accuracy (Smith et al. 2023). Furthermore, Gábor et al. (2023) showed that even models affected by positional uncertainty in species data can be ecologically interpretable. Another study

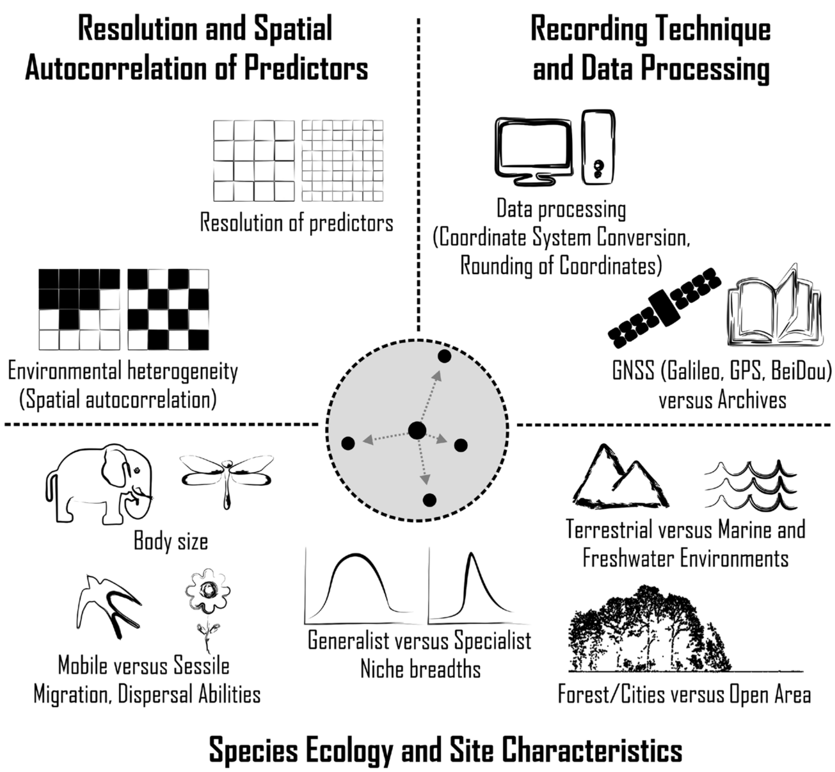

Figure 2. Three groups of interacting factors that determine the magnitude and potential impact of positional uncertainty on species distribution models (SDMs) performance can be specified: the recording technique and data processing (section 'Role of recording technique and data processing'); species ecology and characteristics of the site (section 'Relationships between positional uncertainty, species ecology and ecosystem characteristics'); and the spatial resolution and degree of spatial autocorrelation of the predictors (section 'Relationship between positional uncertainty, spatial resolution and autocorrelation').

Table 2. Studies analysing the influence of positional uncertainty in species occurrence data on species distribution models (SDMs).

Study	Species data	Resolution of environmental var.	Range of shifting occurrences	
			Distance	Cells
Graham et al. 2008	Observed	100 × 100 m	0–5 km	0–50 cells
Johnson and Gillingham 2008	Observed	30 × 30 m	50–1000 m	1–34 cells
Osborne and Leitao 2009	Observed	1 × 1 km	0–5 km	0–5 cells
Fernandez et al. 2009	Observed	1 × 1 km	5–50 km	1–50 cells
Naimi et al. 2011	Virtual	Artificial data	Not valid	1–30 cells
Mitchell et al. 2017	Observed	2.5 × 2.5 m	5–400 m	1–160 cells
Velásquez-Tibatá et al. 2016	Virtual	150 × 150 cells	Not valid	5–15 cells
Gábor et al. 2020b	Virtual	5 × 5 m	5–500 m	1–100 cells
Gábor et al. 2023	Virtual	50 × 50 m	50–1500 m	1–30 cells
Gábor et al. 2023	Observed	200 × 200 m	1–30 km	1–30 cells

investigating the effect of positional uncertainty concluded that models with small sample sizes were more affected by positional uncertainty than models based on larger sample sizes (Mitchell et al. 2017).

The role of positional uncertainty is rarely considered in the evaluation of SDMs. Surprisingly, most SDM studies dealing with positional uncertainty only focused on the training dataset, while ignoring the (potential) effect of inaccurately georeferenced data in the validation dataset. The ultimate consequence of positional uncertainty in species data lies in an erroneous identification of the presence or absence in a given cell (i.e. in specific environmental conditions). In this regard, Foody (2011) demonstrated that validation data should be error-free (i.e. correctly distinguish between presences and absences), as even a small amount of error could result in misidentification of presences/absences and substantial misestimation of model performance. Therefore, data correctly labelled as species presence or absence (i.e. with minimal positional uncertainty) are essential for assessing model performance. More recently, Moudrý et al. (2017) showed that the inclusion of potentially erroneous presences (in this case ambiguous breeding bird categories used in the breeding bird atlases, i.e. possible and probable breeding) severely affected models' performance metrics when added to the validation dataset, while it had a relatively minor effect on model performance when added to the training dataset. Therefore, we suggest relying on large sample size, possibly including observations with low positional accuracy (i.e. with higher positional uncertainty than the spatial resolution of predictors) for model calibration, while preserving high-accuracy data for model validation.

Alternatively, Moudrý and Šímová (2012) suggested that knowing the positional uncertainty of the occurrences allows balancing high- and poor-quality data in both training and testing datasets, e.g. by including a predictor in the model (even as a categorical variable with a few levels of data positional uncertainty) to be tested or to up/downweight the importance of observations (see Velásquez-Tibatá et al. 2016 for such an approach using Bayesian models). This allows preserving most of the data and offsetting the potential negative effect of high positional uncertainty. On the other hand, if the predictor has many levels and few observations (per level), it might be better to subset the data to retain only those of the best quality. If only a small sample size is available, we recommend considering the use of methods to mitigate positional uncertainty (Hefley et al. 2014, Zhang et al. 2018, Smith et al. 2023). Note, however, that the existing approaches typically either require knowledge of the magnitude of the uncertainty and that their use is limited to data with relatively small positional uncertainty (Zhang et al. 2018), or they require that at least part of the dataset is recorded with minimal positional uncertainty (Hefley et al. 2014, Smith et al. 2023). Although recent literature is favouring the inclusion of observations with reasonable positional uncertainty rather than reducing sample size (Gábor et al. 2023, Smith et al. 2023), we recommend careful consideration of this trade-off. Whether it is preferable to maintain the sample size or to minimise the adverse effect of positional uncertainty remains a very timely and unanswered question.

Role of recording technique and data processing

Old datasets, such as historical observations archived in museums, atlases, and natural history collections that were retrospectively georeferenced, are usually thought to be more prone to relatively higher positional error than new ones (Graham et al. 2004, Wieczorek et al. 2004, Newbold 2010, Bloom et al. 2018, Marcer et al. 2022). However, positional error affects any dataset, including those georeferenced using modern technologies such as the global navigation satellite systems (GNSS). Indeed, several factors can degrade GNSS positional accuracy, including the number and position of satellites, and the characteristics of the study site (e.g. beneath a dense forest canopy versus an open grassland). The use of a low number of satellites to georeference species data may be due to the use of outdated technology, such as the use of a device that relies only on the US Global Positioning System (GPS), instead of using all currently available systems (e.g. Galileo, Glonass, and Beidou). Even when the above-mentioned challenges are overcome, species occurrence data may still be impacted by errors introduced during data processing (e.g. wrong transformations among coordinate reference systems, rounding of coordinates, or lack of error correction procedures such as post-differential correction; Sillero and Gonçalves-Seco 2014). Unfortunately, the positional uncertainty of species records is often undocumented (Moudrý and Devillers 2020, Marcer et al. 2022).

Relationships between positional uncertainty, species ecology, and ecosystem characteristics

It is usually impossible to accurately georeference positions for non-sessile species (unless they are equipped with transmitters) due to environmental barriers (for example, it is impossible to get close to the species in some habitats) and/or species characteristics (e.g. size, mobility, and behaviour) (Frair et al. 2010). Besides, georeferencing species' location using GNSS in a dense forest or at the bottom of a narrow and deep ravine may be difficult due to the poor reception of the satellite signal. In addition, buildings, walls, and trees in the proximity of an antenna can reflect the signal from satellites, thereby further reducing the positioning accuracy (a phenomenon known as multipath; Kos et al. 2010). Besides, GNSS does not work underwater; in effect, the positioning of species observations in marine and freshwater environments is based on acoustic positioning, which leads to a decrease in accuracy with the water column depth, or simply on recording a position at the surface of water and disregarding movements of the sampling gear in the water column (Rattray et al. 2014, Mitchell et al. 2017). As a result, data for mobile animals can have a positional uncertainty of tens to hundreds of metres. The distance between an animal and the observer is positively associated with the species' body size and, therefore, big animals are typically less accurately georeferenced as they move a lot or can be even dangerous, which leads to recording their location from a distance (Zhang et al. 2018).

The effect of positional uncertainty on SDMs may also depend on the species' mobility, expressed as the daily dispersal range or migration ability. Many birds, fishes, and big predators are very mobile, and the accurate georeferencing of their location may play a smaller role in SDM performance than in the case of sessile species (see Fig. 2 for an overview of the factors that may interact with the magnitude of positional uncertainty when building SDMs). In this regard, Gábor et al. (2023) showed that the performance of a band-tailed pigeon SDM only slightly decreased with increasing positional uncertainty, while virtual species simulations that did not consider species mobility showed a rapid decrease in SDM performance. Although positional uncertainty seems to depend on species characteristics, its role in affecting SDMs for different groups (such as insects versus big mammals; mobile organisms like birds versus sessile organisms like plants, corals, etc.) is understudied. Among the few studies that analysed the interaction between positional uncertainty and species ecology, Velásquez-Tibatá et al. (2016) and, more recently, Gábor et al. (2020b), showed that positional uncertainty has a greater impact on SDMs' performances for specialists (i.e. species with a narrow niche breadth) than for generalist species (i.e. those with a wide niche breadth). This is due to occurrences of specialist species being more susceptible to a shift into unsuitable environments.

Relationships between positional uncertainty, spatial resolution, and autocorrelation

The spatial resolution of predictors used in SDMs is another critical factor determining the impact of positional uncertainty on model performance. Previous studies on positional uncertainty considered shifts from 5 m up to 50 km. Such a range of uncertainty results in a less impactful shift of species data over raster cells (and across environmental conditions) in a coarse-resolution set of environmental layers (e.g. 1×1 km) than in a fine-resolution set of environmental layers (e.g. 10×10 m). Note that more recent studies investigated shifts of the species occurrence data by up to 160 pixels (which is almost threefold compared to older studies) thanks to the reduced pixel sizes in the current environmental data (see Table 2 for the combinations of adopted resolution and positional uncertainty in existing studies). Indeed, with today's availability of high spatial resolution predictors, misuse of positionally inaccurate species occurrences is increasingly likely, with the risk of exacerbating the negative effect of positional uncertainty on SDMs' performances.

To reduce the effect of positional uncertainty, multiple studies suggested adjusting spatial resolution so that the largest positional uncertainty associated with occurrence data is lower than the spatial resolution of the predictors (Engler et al. 2004, Moudrý and Šímová 2012, Keil et al. 2014, Vollering et al. 2016, Sillero et al. 2021a). However, coarsening the spatial resolution of the environmental variables may degrade information on fine-scale heterogeneity in environmental variables, eventually reducing their explanatory power for predicting species distribution (Mertes and Jetz 2018). In addition, spatial resolution can be coarsened to a level that is too far from the relevant ecological scale (Lecours et al. 2015, Moudrý et al. 2023). Recently, Gábor et al. (2022) showed that coarsening the spatial resolution to compensate for positional uncertainty does not improve model performance. However, they used a relatively simple virtual species approach, so more studies, preferably using 'real' species, are needed to validate their results. Whether maintaining the spatial resolution of the response variable close to the ecological scale is more important than minimising the adverse effect of positional uncertainty (or whether the opposite is true) remains a very current and unanswered question (see Moudrý et al. 2023 for a review of practices for appropriate grain selection).

It is crucial to recognise that shifting species records in the geographic space does not necessarily translate to an equivalent shift in the environmental space. High positional uncertainty can lead to mischaracterizing the conditions under which a species occurs, especially in regions characterised by steep ecological gradients, such as mountainous areas or heavily fragmented landscapes. Indeed, the impact of positional uncertainty is related to the spatial autocorrelation of environmental variables. Naimi et al. (2011) found that the impact of positional uncertainty on SDMs' prediction performance decreased with increasing spatial autocorrelation in the environmental variables. In this regard, examining the degree of spatial autocorrelation in environmental variables was suggested as a way to a priori assess the impact of positional uncertainty on SDMs predictions (Naimi et al. 2011, 2014).

Recommendations associated with positional uncertainty

It is crucial to consider data quality and to carefully assess the implications of using data affected by positional uncertainty in either the training or validation process. Such considerations will yield more reliable assessments of model performance and improve the accuracy of SDMs.

- First, we recommend 'cleaning' the dataset and removing aberrant errors (e.g. records with switched latitude and longitude, or records located at zoos or botanical gardens). This can be performed using automated methods such as those implemented by the 'CoordinateCleaner' R package (Zizka et al. 2019).
- Second, researchers should quantify the positional uncertainty of the remaining input data, for example, using attributes specifying positional uncertainty. If such assessment is limited by metadata availability, for example in the case of historical data, it is recommended to at least approximate the positional uncertainty based on known information, such as the collection methodology or the number of decimals recorded with coordinates (Peterson and Samy 2016, Watcharamongkol et al. 2018, Moudrý and Devillers 2020).
- Third, we recommend researchers to carefully weigh the trade-offs between positional uncertainty and spatial resolution of environmental variables, with greater emphasis on the use of a resolution as close to the ecological scale as possible (Gábor et al. 2022, Moudrý et al. 2023). Preferably, the positional uncertainty should be lower than the spatial resolution of the environmental variables (Moudrý and Šímová 2012). We suggest that the spatial resolution should be at least twice the positional uncertainty to reduce the risk of miscalculation of species–environment relationships. However, this may not always be achievable. In such a case, it is important to consider the following steps to estimate and acknowledge the potential impact of positional uncertainty on the performance of the model.
- Fourth, we suggest considering positional uncertainty in light of the particular species' ecology as some groups of species, such as mobile species, might be less affected by positional uncertainty than others (Gábor et al. 2020b).
- Fifth, researchers should examine the spatial autocorrelation in predictors to gain insight into whether predictions are likely to be affected by positional uncertainty (Naimi et al. 2011, 2014). This may include testing the impact of various resolutions on model performance.
- Finally, we recommend considering the use of methods to mitigate positional uncertainty (Hefley et al. 2014, Zhang et al. 2018, Smith et al. 2023). Alternatively, knowing the positional uncertainty of the occurrences allows the inclusion of predictors in the model to be tested or to up/downweight the importance of observations (Moudrý and Šímová 2012, Velásquez-Tibatá et al. 2016). For new surveys, we suggest using measurement techniques that minimise positional uncertainty, such as differential GNSS (Sillero et al. 2021b), and providing an estimate of the measurement accuracy (as is increasingly common in global databases).

Sampling bias

Sampling bias poses a significant challenge in SDMs, leading to models that provide a partial or distorted view of species distribution or ecological niche (Kadmon et al. 2004, Leitão et al. 2011, Bean et al. 2012, Beck et al. 2014, Stolar and Nielsen 2015, Bardon et al. 2021). Despite advances, our knowledge of species distributions still remains limited for most taxa due to the variations in the sampling intensity over time and huge regions of the world remaining poorly sampled (Isaac and Pocock 2015, Menegotto and Rangel 2018, Hughes et al. 2021, Daru and Rodriguez 2023). Typically, positive sampling biases have been reported towards easily accessible areas (e.g. proximity to roads, rivers, and urban settlements, Kadmon et al. 2004), protected areas (Boakes et al. 2010, Girardello et al. 2019), more populated areas (Geldmann et al. 2016), and charismatic species (Troudet et al. 2017), leading to spatial and taxonomic biases (Huges et al. 2021). Uneven data-sharing practices further exacerbate this issue (Meyer et al. 2015). Various methods have been proposed to compensate for sampling bias in species occurrence records, aiming to create models with quality comparable to models developed with unbiased data. Prevalent approaches for bias compensation include adjusting background samples (the target-group background, TGB, approach; Phillips et al. 2009) in presence–background models, or filtering (thinning) presences (Veloz 2009) (Table 3).

The rationale behind the TGB approach is to select background data with the same sampling bias as the set of presence records (i.e. to bias the background locations towards areas where the presences were sampled; Phillips et al. 2009). The TGB approach adjusts the selection of the background data by assessing the 'sampling effort', which indicates the effort invested during sampling. For example, the TGB approach restricts the sampling of background data to locations where other species of the same order or family as the target species have been observed (preferably using the same methodology/database). This is done assuming that hypothetical surveys would have detected the focal species if it had been present in those locations. Therefore it is especially useful for large citizen science projects (Barber et al. 2022, Boyd et al. 2023) but less suitable for poorly sampled regions where information on the target group may not be available. An appropriately selected target-group background leads to a more reliable estimation of species–environment relationships. Note, however, the importance of careful selection of target group species, as the density of occurrences not only reflects sampling effort but also the varied densities of species and their ecological preferences, potentially introducing new biases (Botella et al. 2020).

Table 3. Studies that evaluated the effect of sampling bias and the effectiveness of methods proposed to compensate for sampling bias on model performance. TGB, target-group background.

Study	Number of species	Bias type	Evaluation approach	Bias correction	Main conclusion
Phillips et al. 2009	226	Existing	Independent data	TGB	Bias correction improve models
Bystriakova et al. 2012	5 plants (*Asplenium* spp.)	Existing	Independent data (but only presences)	TGB	Bias correction improve models
Kramer-Schadt et al. 2013	Malay civet, two virtual species	Existing, Simulated	Simulated data	Geographic filtering, TGB	Geographic filter is preferred relative to TGB
Syfert et al. 2013	Tree fern	Existing	Independent data	TGB	Bias correction improve models
Fourcade et al. 2014	Turtle, salamander, virtual species	Simulated	Original model based on unbiased data	Five methods	Variable efficiency, further research needed
Varela et al. 2014	Virtual	Simulated	Original model based on unbiased data	Environmental and geographic filtering	Recommend environmental filtering
Ranc et al. 2017	Virtual	Simulated	True distribution of simulated species	TGB	Bias correction is detrimental for some species
Castellanos et al. 2019	Virtual	Simulated	True distribution of simulated species	Environmental and geographic filtering	Recommend environmental filtering
Gábor et al. 2020a	Virtual	Simulated	True distribution of simulated species	Environmental filtering	Filtering is not necessarily helpful
Chauvier et al. 2021	1,900 plants	Existing	Independent data	Bias covariate correction, and environmental bias correction	Combining both methods might be the best choice
Inman et al. 2021	Virtual	Simulated	True distribution of simulated species	TGB, geographic and environmental filtering	Bias correction is detrimental for some species
Baker et al. 2022	Virtual	Simulated	True distribution of simulated species	Geographic filtering	More mechanistic understanding of how sampling biases arise is needed

The filtering approach (or thinning) was designed to reduce the negative effect of sampling bias by reducing the number of presences in oversampled regions in the geographic space (Veloz 2009) or oversampled environmental conditions in the environmental space (Varela et al. 2014). Both geographic and environmental filtering use a distance between presences to determine the filter size. However, while geographic filtering uses distances in the geographic space (e.g. latitude and longitude), environmental filtering uses the range between values of multiple environmental variables (Varela et al. 2014, Castellanos et al. 2019). Another strategy carried out in the environmental space is to use presence data (i.e. their position in the environmental space) to identify and filter out background points likely associated with suitable habitats (Da Re et al. 2023). Many studies have evaluated the performance of these methods, simulating the bias by sub-sampling the original data (i.e. a presumably complete dataset without any bias) or by addressing bias already present in the datasets (Table 3). Such assessments require independent evaluation data containing both presence and absence records or comparison against models based on the unbiased dataset before sub-sampling simulation.

Should the bias be assessed in the geographic or environmental space?

There is an ongoing debate about whether bias should be assessed in the geographic or environmental space, or both (Varela et al. 2014, Moudrý 2015, Cosentino and Maiorano 2021, Xu et al. 2024). According to Hutchinson's duality, there is a correspondence between the species' niche in environmental space and its distribution in geographic space. This means that the environmental conditions where a species occurs (its ecological niche) are reflected in its geographic distribution. Conversely, the geographic distribution of a species can provide insights into its ecological niche requirements (Colwell and Rangel 2009). In theory, every location in geographic space can be 'uniquely' characterised by the environmental conditions at that location. However, projections of subsets of environmental space into geographic space can have complicated structures (i.e. a single point in environmental space may correspond to many locations in geographic space; see Colwell and Rangel 2009, Soberón and Nakamura 2009). If only partial knowledge of the ecological niche of a species is available, predicting its distribution in geographic space may result in the omission of multiple locations. On the other hand, a missing site in the geographic space may be substituted by another site with the same environmental conditions. Consequently, the challenge in estimating species–environment relationships lies not only in the spatial bias within the geographic space where the bias originates but also in how this bias is reflected in the environmental space (i.e. the ecological niche space). All SDMs are not purely spatial methods (like interpolation, for instance), and the calculations actually occur within the environmental space defining the species' ecological niche. Therefore, addressing bias within the environmental space directly tackles the model calibration.

Sampling bias is influenced by the sampling design (Hirzel and Guisan 2002, Tessarolo et al. 2014, Mateo et al. 2018, Bazzichetto et al. 2023). A fundamental assumption underlying presence–background methods is that environmental conditions are sampled in proportion to their actual availability (Hastie and Fithian 2013). Note that it is not a geographic space where uniform sampling is required but rather the environmental conditions that have to be sampled in proportion to their availability (Aarts et al. 2012, Merow et al. 2013). If this is not fulfilled, clustered occurrences may lead to the overestimation of the environmental suitability for the respective species in environments that have been sampled more intensively (e.g. environments in protected areas, or near roads and towns) and underestimated for those surveyed less intensively (Barry and Elith 2006, Guillera-Arroita et al. 2015). For instance, fully random draws of species' presence in the geographic space may introduce a bias towards the most widespread environmental conditions, which possibly leads to uneven sampling of the species' niche within the environmental space (Bazzichetto et al. 2023). This issue is associated with another underlying assumption: that the species' niche is comprehensively sampled across the entire spectrum of environmental conditions in which it occurs (Phillips et al. 2009). Failing to meet this assumption, which can happen when there is a lack of knowledge about a species' tolerance to abiotic conditions (i.e. environmental bias), may cause a poor estimation of the actual niche occupied by the species (Hortal et al. 2008). If the ecological niche of the species is truncated (i.e. the complete niche of the species is not captured by the occurrences), it is not possible to extrapolate a reliable model into different spatial or temporal dimensions (Chevalier et al. 2022). Therefore, representative sampling of the environmental space should in principle give better results, regardless of its bias in the geographic space (Tessarolo et al. 2014, Sabatini et al. 2021, Bazzichetto et al. 2023).

We recommend considering both geographic and environmental spaces in the assessment of sampling bias (Tessarolo et al. 2014, Cosentino and Maiorano 2021). In areas of high geographic and high environmental bias, and particularly in undersampled environments, further sampling efforts are required. Alternatively, bias correction based on the TGB method or geographic filtering can be suitable options (Inman et al. 2021), although the latter was recently strongly criticised, and its effectiveness in mitigating sampling biases is being questioned (Ten Caten and Dallas 2023, Lamboley and Fourcade 2024). Given that geographic filtering reduces the sample size, TGB seems to be a better alternative (Barber et al. 2022). However, a bias in the geographic space does not necessarily lead to a bias in the environmental space. If the geographic bias is high but the environmental bias is low, no corrections are needed, and the data can be used 'as is' for modelling. For example, Kadmon et al. (2004) and more recently Mccarthy et al. (2012) showed that collecting data close to roads can still provide an adequate sampling of ecological gradients if the road network has high environmental coverage, thus allowing the uncovering of the true species–environment relationships. In the case of low geographic but high environmental bias, further sampling of undersampled environments is preferable; however, if it is not possible, it is reasonable to consider directly a correction in the environmental space using environmental filtering (Varela et al. 2014, Cosentino and Maiorano 2021). Nevertheless, see the risks of performing this procedure described in the following paragraph.

Geographic and environmental spaces are communicating vessels, and so correcting one component (geographic or environmental) may have a detrimental effect on the other. For example, geographical filtering could unwittingly exclude occurrences in the environmental space with unique environmental conditions or disguise true patterns, e.g. due to clustering for ecological reasons such as breeding, social behaviour, or predator–prey dynamics (Varela et al. 2014). On the other hand, environmental filtering (downweighting repeated species occurrences in similar environmental conditions) identifies grid cells within marginal habitats to be equally suitable as the cells representing core habitats. For example, if the species probability of occurrence is 0.1 at one site and 0.7 at another, such sites will be occupied in one and seven out of 10 cases, respectively. If we disregard the presences at the latter site, we lose the ability to discern the conditions favoured by the species (Moudrý et al. 2015). Indeed, it is impossible to use presence–background data to determine whether species observed in particular environments result from a larger sampling effort or ecological preferences (Guillera-Arroita et al. 2015), and removing bias without the information on the sampling effort becomes quixotic (Rocchini et al. 2023).

How sampling bias (and correction methods) interact with species ecology

Several studies have reported that there was no improvement or even detrimental effects on SDMs performance after filtering out biased samples (Chefaoui and Serrão 2017, Ranc et al. 2017, Gábor et al. 2020a), and it has been suggested that this might be related to species ecology (Bystriakova et al. 2012). For example, Ranc et al. (2017) showed that range size was the most important factor driving species vulnerability to sampling bias, and that widespread species were more affected by sampling bias and more likely to benefit from bias correction than species with narrow geographic ranges. Similarly, Baker et al. (2022) showed that species type has a notable effect on model performance, with models generally being more robust to the effects of sampling bias for specialist (narrow environmental niches) than for generalist (wide environmental niches) species. In addition, a few studies highlighted that bias correction was detrimental for species with narrow ranges (Ranc et al. 2017), narrow niches (Inman et al. 2021), or low prevalence (Gábor et al. 2020a) and yielded worse models than without bias correction. It is evident that different species are differently affected by sampling bias and respond differently to bias correction. Therefore, species ecology should be considered when correcting for sampling bias.

Recommendations associated with sampling bias

Complete elimination of spatial bias from the modelling procedure is impossible without proper knowledge of all the processes generating it (Rocchini et al. 2023), and it is unrealistic to assume that sampling bias in biodiversity data can be eliminated, even with the development of automated observation technologies. Hence, SDMs need to explore and acknowledge the inherent biases associated with the data in both the geographic and environmental space (Cosentino and Maiorano 2021, Rocchini et al. 2023).

- First, researchers should quantify the sampling bias of their input data in the geographic space. For example, the 'sampbias' R package (Zizka et al. 2021) can be used for such purposes.
- Second, bias should also be evaluated in the environmental space by comparing the distribution of the cells where the focal species was present to all cells in the study area in a gridded environmental space of ecological predictors. This can be done, for example, by using ecological niche factor analysis (Hirzel et al. 2002); 'hypervolume' R package (Blonder et al. 2014); or principal component analysis in the 'ecospat' R package (Di Cola et al. 2017).
- The relationship between geographic and environmental bias should be further explored using local indicators of spatial association (LISA; Anselin 1995) and the results of such an assessment should be used as a basis for the selection of bias-correction methods (Cosentino and Maiorano 2021, Rocchini et al. 2023). This quantification can also assist researchers in effectively directing their additional sampling efforts.
- The next step lies in the application of the bias-correction method, if necessary. Filtering or the TGB approach are possible options, but caution is needed as they could result in lower model performance in particular cases. This requires consideration of species' ecology, as specialist species typically do not benefit from bias correction or can even be negatively affected by it (Gábor et al. 2020a, Inman et al. 2021, Baker et al. 2022). In addition, it is important to notice that filtering will inevitably reduce the number of presences available for modelling. Therefore, if the sample size is relatively small, the TGB approach might be a preferred method (or alternatives such as that proposed by Da Re et al. 2023 for filtering background points implemented in the 'USE' R package).

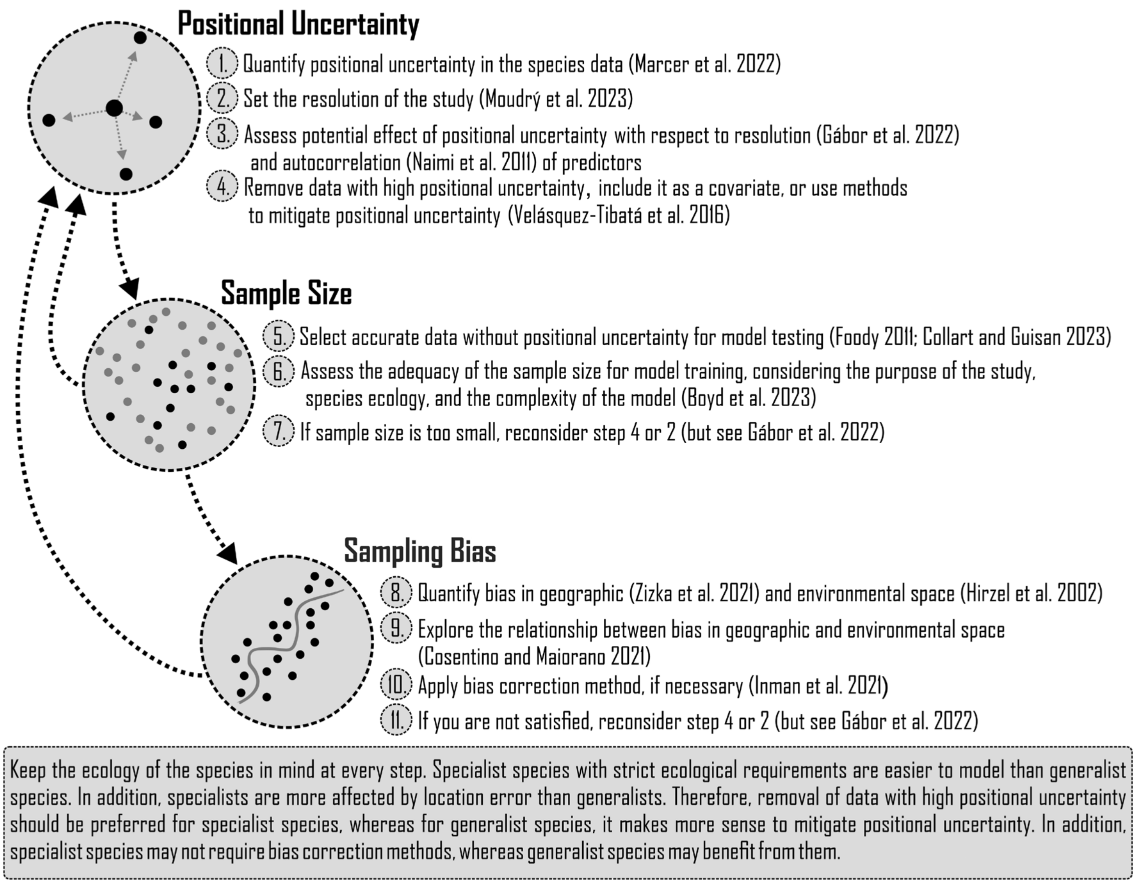

Figure 3. Workflow for a critical assessment of spatial data to be used in species distribution models (SDMs). For more information on the individual steps, see the 'Recommendations' subsections at the end of each main section.

Guidelines and future directions

Despite the increasing number of studies focusing on how various limitations inherent to species data affect the performance of SDMs, there are still gaps in our knowledge, and the use of SDMs remains problematic in many contexts. To advance our understanding, future studies should focus on comprehensive analyses that simultaneously consider various issues, such as sample size, sampling bias (in the geographic and environmental space), positional uncertainty, spatial resolution, and the interaction between the former factors and species' ecological characteristics (Fig. 1). Such studies can help establish the urgently needed guidelines for better-informed modelling choices (e.g. bias correction, removal of data with high positional uncertainty and its effect on sample size and SDMs performance) concerning data limitations and species ecology. Regarding species characteristics, it is important to do such evaluations on characteristics that are easy to specify (i.e. we know them for the majority of species), such as species' niche breadth (generalist versus specialist species), dispersal ability, body size, or trophic group. This way, the assessments can be further used to guide data selection processes in other studies. The consideration of data limitations is crucial in every domain where SDMs are used (Araújo and Peterson 2012, Guisan et al. 2013). These include the discovery of new populations (Fois et al. 2015), reserve selection and design (Esselman and Allan 2011), species translocations or reintroductions (Segal et al. 2021), biological invasions and disease transmission studies (Peterson 2014, Peterson and Samy 2016, Johnson et al. 2019), investigations of climate change impacts (Ehrlén and Morris 2015, Haesen et al. 2023), or testing of biogeographical or evolutionary hypotheses (Machado et al. 2019).

Finally, it is crucial to transparently report bias and uncertainty in the data used for modelling. This includes quantifying sampling bias in geographical and environmental space, as well as positional uncertainty concerning the spatial resolution and autocorrelation of predictors (Fig. 3). Reporting on how species occurrences were divided into training and testing datasets, whether their positional uncertainty was considered and, if applicable, which ones were removed and what was the impact on sample size. Whenever possible, rigorous tests should be conducted to examine the impact of geographical and environmental bias, as well as of positional uncertainty, on model performance (e.g. indicating which approaches were considered to minimise bias and positional uncertainty and their results). Until more comprehensive assessments are available, we strongly recommend remaining vigilant about data limitations and following the basic guidelines for a critical assessment of spatial data to be used in SDMs shown in Fig. 3. The data collection methods, preprocessing, model fitting, and quality assessments, can be reported using standard protocol for reporting SDMs' overview, data, model, assessment, and prediction (ODMAP; Zurell et al. 2020).

Funding – Funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. This work was funded by the Horizon Europe project EarthBridge (grant agreement no. 101079310). RR and AFC were supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2070 – 390732324. MB acknowledges funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 101066324. RGM was funded by project grants Connect2restore (TED2021-129589B-I00, funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR), and NextDive (PID2021-124187NB-I00, funded by MCIN/AEI/10.13039/501100011033 and by ERDF, a way of making Europe). AZ-A was supported by a Margarita Salas Contract financed by the European Union-NextGenerationEU, Ministerio de Universidades y Plan de Recuperacion, Tranformacion y Resiliencia, Spain. MJM, JW, and JP were supported by the Czech Academy of Sciences (project RVO 67985939). FL was funded by the European Union (ERC, BEAST, 101044740). JJL was supported by BiodivERsA+ (ASICS project (G0H6720N, BiodivClim call 2019-2020)). NS was supported by a CEEC2017 contract (CEECIND/02213/2017) from FCT – Fundação para a Ciência e a Tecnologia, Portugal. MT was partially funded by the European Union's Horizon 2020 research and innovation program under grant agreement no. 862480 (SHOWCASE).

Author contributions

Vítězslav Moudrý: Conceptualization (lead); Visualization (equal); Writing – original draft (lead); Writing – review and editing (equal). **Manuele Bazzichetto**: Conceptualization (equal); Writing – review and editing (equal). **Ruben Remelgado**: Conceptualization (equal); Writing – review and editing (equal). **Rodolphe Devillers**: Conceptualization (equal); Writing – review and editing (equal). **Jonathan Lenoir**: Conceptualization (equal); Writing – review and editing (equal). **Rubén G. Mateo**: Conceptualization (equal); Writing – review and editing (equal). **Jonas J. Lembrechts**: Conceptualization (equal); Writing – review and editing (equal). **Neftalí Sillero**: Conceptualization (equal); Writing – review and editing (equal). **Vincent Lecours**: Conceptualization (equal); Writing – review and editing (equal). **Anna F. Cord**: Conceptualization (equal); Writing – review and editing (equal). **Vojtěch Barták**: Conceptualization (equal); Writing – review and editing (equal). **Petr Balej**: Conceptualization (equal); Writing – review and editing (equal). **Duccio Rocchini**: Conceptualization (equal); Writing – review and editing (equal). **Michele Torresani**: Conceptualization (equal); Writing – review and editing (equal). **Salvador Arenas-Castro**: Conceptualization (equal); Writing – review and editing (equal). **Matěj Man**: Conceptualization (equal); Writing – review and editing (equal). **Dominika Prajzlerová**: Conceptualization (equal); Writing – review and editing (equal). **Kateřina Gdulová**: Conceptualization (equal); Writing – review and editing (equal). **Jiří Prošek**: Visualization (equal); Writing – review and editing (equal); **Elisa Marchetto**: Conceptualization (equal); Writing – review and editing (equal). **Alejandra Zarzo-Arias**: Conceptualization (equal); Writing – review and editing (equal). **Lukáš Gábor**: Conceptualization (equal); Writing – review and editing (equal). **François Leroy**: Conceptualization (equal); Writing – review and editing (equal). **Matilde Martini**: Conceptualization (equal); Writing – review and editing (equal). **Marco Malavasi**: Conceptualization (equal); Writing – review and editing (equal). **Roberto Cazzolla Gatti**: Conceptualization (equal); Writing – review and editing (equal). **Jan Wild**: Conceptualization (equal); Writing – review and editing (equal). **Petra Šímová**: Conceptualization (equal); Writing – review and editing (equal).

Transparent peer review

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/ecog.07294.

Data availability statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

Aarts, G., Fieberg, J. and Matthiopoulos, J. 2012. Comparative interpretation of count, presence–absence and point methods for species distribution models. – Methods Ecol. Evol. 3: 177–187.

Anselin, L. 1995. Local indicators of spatial association – LISA. – Geogr. Anal. 27: 93–115.

Araújo, M. B. and Peterson, A. T. 2012. Uses and misuses of bioclimatic envelope modeling. – Ecology 93: 1527–1539.

Araújo, M. B., Anderson, R. P., Márcia Barbosa, A., Beale, C. M., Dormann, C. F., Early, R., Garcia, R. A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R. B., Zimmermann, N. E. and Rahbek, C. 2019. Standards for distribution models in biodiversity assessments. – Sci. Adv. 5: eaat4858.

Arenas-Castro, S., Regos, A., Martins, I., Honrado, J. and Alonso, J. 2022. Effects of input data sources on species distribution model predictions across species with different distributional ranges. – J. Biogeogr. 49: 1299–1312.

Austin, M. P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. – Ecol. Modell. 157: 101–118.

Baartman, J. E., Melsen, L. A., Moore, D. and van der Ploeg, M. J. 2020. On the complexity of model complexity: viewpoints across the geosciences. – Catena 186: 104261.

Baker, D. J., Maclean, I. M., Goodall, M. and Gaston, K. J. 2022. Correlations between spatial sampling biases and environmental niches affect species distribution models. – Global Ecol. Biogeogr. 31: 1038–1050.

Barber, R. A., Ball, S. G., Morris, R. K. and Gilbert, F. 2022. Target-group backgrounds prove effective at correcting sampling bias in Maxent models. – Divers. Distrib. 28: 128–141.

Bardon, L. R., Ward, B. A., Dutkiewicz, S. and Cael, B. B. 2021. Testing the skill of a species distribution model using a 21st century virtual ecosystem. – Geophys. Res. Lett. 48: e2021.

Barry, S. and Elith, J. 2006. Error and uncertainty in habitat models. – J. Appl. Ecol. 43: 413–423.

Bazzichetto, M., Massol, F., Carboni, M., Lenoir, J., Lembrechts, J. J., Joly, R. and Renault, D. 2021. Once upon a time in the far south: influence of local drivers and functional traits on plant invasion in the harsh sub-Antarctic islands. – J. Veg. Sci. 32: e13057.

Bazzichetto, M., Lenoir, J., Da Re, D., Tordoni, E., Rocchini, D., Malavasi, M., Barták, V. and Sperandii, M. G. 2023. Sampling strategy matters to accurately estimate response curves' parameters in species distribution models. – Global Ecol. Biogeogr. 32: 1717–1729.

Bean, W. T., Stafford, R. and Brashares, J. S. 2012. The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. – Ecography 35: 250–258.

Beck, J., Böller, M., Erhardt, A. and Schwanghart, W. 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. – Ecol. Inform. 19: 10–15.

Becker, F. S., Slingsby, J. A., Measey, J., Tolley, K. A. and Altwegg, R. 2022. Finding rare species and estimating the probability that all occupied sites have been found. – Ecol. Appl. 32: e2502.

Bell, D. M. and Schlaepfer, D. R. 2016. On the dangers of model complexity without ecological justification in species distribution modeling. – Ecol. Modell. 330: 50–59.

Blonder, B., Lamanna, C., Violle, C. and Enquist, B. J. 2014. The n-dimensional hypervolume. – Global Ecol. Biogeogr. 23: 595–609.

Bloom, T. D. S., Flower, A. and DeChaine, E. G. 2018. Why georeferencing matters: introducing a practical protocol to prepare species occurrence records for spatial analysis. – Ecol. Evol. 8: 765–777.

Boakes, E. H., McGowan, P. J., Fuller, R. A., Chang-qing, D., Clark, N. E., O'Connor, K. and Mace, G. M. 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. – PLoS Biol. 8: e1000385.

Botella, C., Joly, A., Monestiez, P., Bonnet, P. and Munoz, F. 2020. Bias in presence-only niche models related to sampling effort and species niches: lessons for background point selection. – PLoS One 15: e0232078.

Botella, C., Deneu, B., Marcos, D., Servajean, M., Estopinan, J., Larcher, T., and Joly, A. 2023. The GeoLifeCLEF 2023 dataset to evaluate plant species distribution models at high spatial resolution across Europe. – arXiv preprint arXiv:2308.05121.

Boyd, R. J., Harvey, M., Roy, D. B., Barber, T., Haysom, K. A., Macadam, C. R., Morris, R. K. A., Palmer, C., Palmer, S., Preston, C. D., Taylor, P., Ward, R., Ball, S. G. and Pescott, O. L. 2023. Causal inference and large-scale expert validation shed light on the drivers of SDM accuracy and variance. – Divers. Distrib. 29: 774–784.

Brun, P., Thuiller, W., Chauvier, Y., Pellissier, L., Wüest, R. O., Wang, Z. and Zimmermann, N. E. 2020. Model complexity affects species distribution projections under climate change. – J. Biogeogr. 47: 130–142.

Bystriakova, N., Peregrym, M., Erkens, R. H. J., Bezsmertna, O. and Schneider, H. 2012. Sampling bias in geographic and environmental space and its effect on the predictive power of species distribution models. – Syst. Biodivers. 10: 305–315.

Carretero, M. A. and Sillero, N. 2016. Evaluating how species niche modelling is affected by partial distributions with an empirical case. – Acta Oecol. 77: 207–216.

Castellanos, A. A., Huntley, J. W., Voelker, G. and Lawing, A. M. 2019. Environmental filtering improves ecological niche models across multiple scales. – Methods Ecol. Evol. 10: 481–492.

Chauvier, Y., Zimmermann, N. E., Poggiato, G., Bystrova, D., Brun, P. and Thuiller, W. 2021. Novel methods to correct for observer and sampling bias in presence-only species distribution models. – Global Ecol. Biogeogr. 30: 2312–2325.

Chefaoui, R. M. and Serrão, E. A. 2017. Accounting for uncertainty in predictions of a marine species: integrating population genetics to verify past distributions. – Ecol. Modell. 359: 229–239.

Chevalier, M., Zarzo-Arias, A., Guélat, J., Mateo, R. G. and Guisan, A. 2022. Accounting for niche truncation to improve spatial and temporal predictions of species distributions. – Front. Ecol. Evol. 10: 944116.

Collart, F. and Guisan, A. 2023. Small to train, small to test: dealing with low sample size in model evaluation. – Ecol. Inform. 75: 102106.

Collart, F., Broennimann, O., Guisan, A. and Vanderpoorten, A. 2023. Ecological and biological indicators of the accuracy of species distribution models: lessons from European bryophytes. – Ecography 23: e06721.

Colwell, R. K. and Rangel, T. F. 2009. Hutchinson's duality: the once and future niche. – Proc. Natl Acad. Sci. USA 106: 19651–19658.

Cosentino, F. and Maiorano, L. 2021. Is geographic sampling bias representative of environmental space? – Ecol. Inform. 64: 101369.

Coudun, C. and Gégout, J. C. 2006. The derivation of species response curves with Gaussian logistic regression is sensitive to sampling intensity and curve characteristics. – Ecol. Modell. 199: 164–175.

Da Re, D., Tordoni, E., Lenoir, J., Vanwambeke, S. O., Rocchini, D., Bazzichetto, M. and SoilTemp Consortium. 2023. Use it: uniformly sampling pseudo-absences within the environmental space for applications in habitat suitability models.

Daru, B. H. and Rodriguez, J. 2023. Mass production of unvouchered records fails to represent global biodiversity patterns. – Nat. Ecol. Evol. 7: 816–831.

Davies, S. C., Thompson, P. L., Gomez, C., Nephin, J., Knudby, A., Park, A. E., Friesen, S. K., Pollock, L. J., Rubidge, E. M., Anderson, S. C., Iacarella, J. C., Lyons, D. A., MacDonald, A., McMillan, A., Ward, E. J., Holdsworth, A. M., Swart, N., Price, J. and Hunter, K. L. 2023. Addressing uncertainty when projecting marine species' distributions under climate change. – Ecography2023: e06731.

Di Cola, V., Broennimann, O., Petitpierre, B., Breiner, F. T., d'Amen, M., Randin, C., Engler, R., Pottier, J., Pio, D., Dubuis, A., Pellissier, L., Mateo, R. G., Hordijk, W., Salamin, N. and Guisan, A. 2017. ecospat: an R package to support spatial analyses and modeling of species niches and distributions. – Ecography 40: 774–787.

Duputié, A., Zimmermann, N. E. and Chuine, I. 2014. Where are the wild things? Why we need better data on species distribution. – Global Ecol. Biogeogr. 23: 457–467.

Ehrlén, J. and Morris, W. F. 2015. Predicting changes in the distribution and abundance of species under environmental change. – Ecol. Lett. 18: 303–314.

Elith, J. and Leathwick, J. R. 2009. Species distribution models: ecological explanation and prediction across space and time. – Annu. Rev. Ecol. Evol. Syst. 40: 677–697.

Elith, J., Burgman, M. A. and Regan, H. M. 2002. Mapping epistemic uncertainties and vague concepts in predictions of species distribution. – Ecol. Modell. 157: 313–329.

Engler, R., Guisan, A. and Rechsteiner, L. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. – J. Appl. Ecol. 41: 263–274.

Esselman, P. C. and Allan, J. D. 2011. Application of species distribution models and conservation planning software to the design of a reserve network for the riverine fishes of northeastern Mesoamerica. – Freshw. Biol. 56: 71–88.

Feeley, K. J. and Silman, M. R. 2011. Keep collecting: accurate species distribution modelling requires more collections than previously thought. – Divers. Distrib. 17: 1132–1140.

Feng, X., Park, D. S., Walker, C., Peterson, A. T., Merow, C. and Papeş, M. 2019. A checklist for maximizing reproducibility of ecological niche models. – Nat. Ecol. Evol. 3: 1382–1395.

Fernandez, M., Blum, S., Reichle, S., Guo, Q., Holzman, B. and Hamilton, H. 2009. Locality uncertainty and the differential performance of four common niche-based modeling techniques. – Biodivers. Inform. 6: 36–52.

Ferrier, S., Jetz, W. and Scharlemann, J. 2017. Biodiversity modelling as part of an observation system. The GEO handbook on biodiversity observation networks. – Springer, pp. 239–257.

Ficetola, G. F., Bonardi, A., Mücher, C. A., Gilissen, N. L. M. and Padoa-Schioppa, E. 2014. How many predictors in species distribution models at the landscape scale? Land use versus LiDAR-derived canopy height. – Int. J. Geogr. Inf. Sci. 28: 1723–1739.

Fois, M., Fenu, G., Cuena Lombraña, A. C., Cogoni, D. and Bacchetta, G. 2015. A practical method to speed up the discovery of unknown populations using species distribution models. – J. Nat. Conserv. 24: 42–48.

Fois, M., Cuena-Lombraña, A., Fenu, G. and Bacchetta, G. 2018. Using species distribution models at local scale to guide the search of poorly known species: review, methodological issues and future directions. – Ecol. Modell. 385: 124–132.

Foody, G. M. 2011. Impacts of imperfect reference data on the apparent accuracy of species presence–absence models and their predictions. – Global Ecol. Biogeogr. 20: 498–508.

Fourcade, Y., Engler, J. O., Rödder, D. and Secondi, J. 2014. Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. – PLoS One 9: e97122.

Fourcade, Y., Besnard, A. G. and Secondi, J. 2018. Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. – Global Ecol. Biogeogr. 27: 245–256.

Frair, J. L., Fieberg, J., Hebblewhite, M., Cagnacci, F., DeCesare, N. J. and Pedrotti, L. 2010. Resolving issues of imprecise and habitat-biased locations in ecological analyses using GPS telemetry data. – Phil. Trans. R. Soc. B 365: 2187–2200.

Gábor, L., Moudrý, V., Barták, V. and Lecours, V. 2020a. How do species and data characteristics affect species distribution models and when to use environmental filtering? – Int. J. Geogr. Inf. Sci. 34: 1567–1584.

Gábor, L., Moudrý, V., Lecours, V., Malavasi, M., Barták, V., Fogl, M., Šímová, P., Rocchini, D. and Václavík, T. 2020b. The effect of positional error on fine scale species distribution models increases for specialist species. – Ecography 43: 256–269.

Gábor, L., Jetz, W., Lu, M., Rocchini, D., Cord, A., Malavasi, M., Zarzo-Arias, A., Barták, V. and Moudrý, V. 2022. Positional errors in species distribution modelling are not overcome by the coarser grains of analysis. – Methods Ecol. Evol. 13: 2289–2302.

Gábor, L., Jetz, W., Zarzo-Arias, A., Winner, K., Yanco, S., Pinkert, S., Marsh, C. J., Rogan, M. S., Mäkinen, J., Rocchini, D., Barták, V., Malavasi, M., Balej, P. and Moudrý, V. 2023. Species distribution models affected by positional uncertainty in species occurrences can still be ecologically interpretable. – Ecography 2023: e06358.

Gábor, L., Cohen, J., Moudrý, V. and Jetz, W. 2024. Assessing the applicability of binary land-cover variables to species distribution models across multiple grains. – Landscape Ecol. 39: 66.

García-Callejas, D. and Araújo, M. B. 2016. The effects of model and data complexity on predictions from species distributions models. – Ecol. Modell. 326: 4–12.

Geldmann, J., Heilmann-Clausen, J., Holm, T. E., Levinsky, I., Markussen, B. O., Olsen, K., Rahbek, C. and Tøttrup, A. P. 2016. What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. – Divers. Distrib. 22: 1139–1149.

Girardello, M., Chapman, A., Dennis, R., Kaila, L., Borges, P. A. and Santangeli, A. 2019. Gaps in butterfly inventory data: a global analysis. – Biol. Conserv. 236: 289–295.

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C. and Peterson, A. T. 2004. New developments in museum-based informatics and applications in biodiversity analysis. – Trends Ecol. Evol. 19: 497–503.

Graham, C. H., Elith, J., Hijmans, R. J., Guisan, A., Townsend Peterson, A., Loiselle, B. A. and NCEAS Predicting Species Distributions Working Group. 2008. The influence of spatial errors in species occurrence data used in distribution models. – J. Appl. Ecol. 45: 239–247.

Guillera-Arroita, G., Lahoz-Monfort, J. J., Elith, J., Gordon, A., Kujala, H., Lentini, P. E., McCarthy, M. A., Tingley, R. and Wintle, B. A. 2015. Is my species distribution model fit for purpose? Matching data and models to applications. – Global Ecol. Biogeogr. 24: 276–292.

Guisan, A., Zimmermann, N. E., Elith, J., Graham, C. H., Phillips, S. and Peterson, A. T. 2007. What matters for predicting the occurrences of trees: techniques, data, or species' characteristics? – Ecol. Monogr. 77: 615–630.

Guisan, A. et al. 2013. Predicting species distributions for conservation decisions. – Ecol. Lett. 16: 1424–1435.

Haesen, S., Lenoir, J., Gril, E., De Frenne, P., Lembrechts, J. J., Kopecký, M., Macek, M., Man, M., Wild, J. and Van Meerbeek, K. 2023. Microclimate reveals the true thermal niche of forest plant species. – Ecol. Lett. 26: 2043–2055.

Hallman, T. A. and Robinson, W. D. 2020. Deciphering ecology from statistical artefacts: competing influence of sample size, prevalence and habitat specialization on species distribution models and how small evaluation datasets can inflate metrics of performance. – Divers. Distrib. 26: 315–328.

Hanberry, B. B., He, H. S. and Dey, D. C. 2012. Sample sizes and model comparison metrics for species distribution models. – Ecol. Modell. 227: 29–33.

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

Hefley, T. J., Baasch, D. M., Tyre, A. J. and Blankenship, E. E. 2014. Correction of location errors for presence-only species distribution models. – Methods Ecol. Evol. 5: 207–214.

Heikkinen, R. K., Luoto, M., Araújo, M. B., Virkkala, R., Thuiller, W. and Sykes, M. T. 2006. Methods and uncertainties in bioclimatic envelope modelling under climate change. – Prog. Phys. Geogr. 30: 751–777.

Hernandez, P. A., Graham, C. H., Master, L. L. and Albert, D. L. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. – Ecography 29: 773–785.

Hirzel, A. and Guisan, A. 2002. Which is the optimal sampling strategy for habitat suitability modelling. – Ecol. Modell. 157: 331–341.

Hirzel, A. H., Hausser, J., Chessel, D. and Perrin, N. 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? – Ecology 83: 2027–2036.

Hortal, J., Jiménez-Valverde, A., Gómez, J. F., Lobo, J. M. and Baselga, A. 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. – Oikos 117: 847–858.

Hortal, J., de Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M. and Ladle, R. J. 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. – Annu. Rev. Ecol. Evol. Syst. 46: 523–549.

Hughes, A., Dorey, J., Bossert, S., Qiao, H. and Orr, M. 2023. Big data – big problems? How to circumvent problems in biodiversity mapping and ensure meaningful results. – Ecography 2024: e07115.

Hughes, A. C., Orr, M. C., Ma, K., Costello, M. J., Waller, J., Provoost, P., Yang, Q., Zhu, C. and Qiao, H. 2021. Sampling biases shape our view of the natural world. – Ecography 44: 1259–1269.

Inman, R., Franklin, J., Esque, T. and Nussear, K. 2021. Comparing sample bias correction methods for species distribution modeling using virtual species. – Ecosphere 12: e03422.

Isaac, N. J. and Pocock, M. J. 2015. Bias and information in biological records. – Biol. J. Linn. Soc. 115: 522–531.

Jansen, J., Woolley, S. N., Dunstan, P. K., Foster, S. D., Hill, N. A., Haward, M. and Johnson, C. R. 2022. Stop ignoring map uncertainty in biodiversity science and conservation policy. – Nat. Ecol. Evol. 6: 828–829.

Jeliazkov, A., Gavish, Y., Marsh, C. J., Geschke, J., Brummitt, N., Rocchini, D., Haase, P., Kunin, W. E. and Henle, K. 2022. Sampling and modelling rare species: conceptual guidelines for the neglected majority. – Global Change Biol. 28: 3754–3777.

Jiménez-Valverde, A. 2020. Sample size for the evaluation of presence-absence models. – Ecol. Indic. 114: 106289.

Jiménez-Valverde, A., Lobo, J. and Hortal, J. 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. – Commun. Ecol. 10: 196–205.

Johnson, C. J. and Gillingham, M. P. 2008. Sensitivity of species-distribution models to error, bias, and model design: an application to resource selection functions for woodland caribou. – Ecol. Modell. 213: 143–155.

Johnson, E. E., Escobar, L. E. and Zambrana-Torrelio, C. 2019. An ecological framework for modeling the geography of disease transmission. – Trends Ecol. Evol. 34: 655–668.

Kadmon, R., Farber, O. and Danin, A. 2003. A systematic analysis of factors affecting the performance of climatic envelope models. – Ecol. Appl. 13: 853–867.

Kadmon, R., Farber, O. and Danin, A. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. – Ecol. Appl. 14: 401–413.

Keil, P., Wilson, A. M. and Jetz, W. 2014. Uncertainty, priors, autocorrelation and disparate data in downscaling of species distributions. – Divers. Distrib. 20: 797–812.

Kos, T., Markezic, I. and Pokrajcic, J. 2010. Effects of multipath reception on GPS positioning performance. – In: Grgić, M., Božek, J. and Grgić, S. (eds), Proceedings ELMAR-2010. IEEE, pp. 399–402.

Kramer-Schadt, S. et al. 2013. The importance of correcting for sampling bias in MaxEnt species distribution models. – Divers. Distrib. 19: 1366–1379.
Lamboley, Q. and Fourcade, Y. 2024. No optimal spatial filtering distance for mitigating sampling bias in ecological niche models. – J. Biogeogr., doi: 10.1111/jbi.14854.
Lecours, V., Devillers, R., Schneider, D. C., Lucieer, V. L., Brown, C. J. and Edinger, E. N. 2015. Spatial scale and geographic context in benthic habitat mapping: review and future directions. – Mar. Ecol. Prog. Ser. 535: 259–284.
Leitão, P. J., Moreira, F. and Osborne, P. E. 2011. Effects of geographical data sampling bias on habitat models of species distributions: a case study with steppe birds in southern Portugal. – Int. J. Geogr. Inf. Sci. 25: 439–454.
Liu, C., Newell, G. and White, M. 2019. The effect of sample size on the accuracy of species distribution models: considering both presences and pseudo-absences or background sites. – Ecography 42: 535–548.
Loiselle, B. A., Jørgensen, P. M., Consiglio, T., Jiménez, I., Blake, J. G., Lohmann, L. G. and Montiel, O. M. 2008. Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? – J. Biogeogr. 35: 105–116.
Machado, A. F., Nunes, M. S., Silva, C. R., Dos Santos, M. A., Farias, I. P., da Silva, M. N. F. and Anciães, M. 2019. Integrating phylogeography and ecological niche modelling to test diversification hypotheses using a Neotropical rodent. – Evol. Ecol. 33: 111–148.
Maggini, R., Lehmann, A., Zimmermann, N. E. and Guisan, A. 2006. Improving generalized regression analysis for the spatial prediction of forest communities. – J. Biogeogr. 33: 1729–1749.
Marcer, A., Chapman, A. D., Wieczorek, J. R., Xavier Picó, F., Uribe, F., Waller, J. and Ariño, A. H. 2022. Uncertainty matters: ascertaining where specimens in natural history collections come from and its implications for predicting species distributions. – Ecography 2022: e06025.
Mateo, R. G., Felicísimo, Á. M. and Muñoz, J. 2010. Effects of the number of presences on reliability and stability of MARS species distribution models: the importance of regional niche variation and ecological heterogeneity. – J. Veg. Sci. 21: 908–922.
Mateo, R. G., Gastón, A., Aroca-Fernández, M. J., Saura, S. and García-Viñas, J. I. 2018. Optimization of forest sampling strategies for woody plant species distribution modelling at the landscape scale. – For. Ecol. Manage. 410: 104–113.
McCarthy, K. P., Fletcher Jr, R. J., Rota, C. T. and Hutto, R. L. 2012. Predicting species distributions from samples collected along roadsides. – Conserv. Biol. 26: 68–77.
McPherson, J. M. and Jetz, W. 2007. Effects of species' ecology on the accuracy of distribution models. – Ecography 30: 135–151.
McPherson, J. M., Jetz, W. and Rogers, D. J. 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? – J. Appl. Ecol. 41: 811–823.
Menegotto, A. and Rangel, T. F. 2018. Mapping knowledge gaps in marine diversity reveals a latitudinal gradient of missing species richness. – Nat. Commun. 9: 4713.
Merow, C., Smith, M. J. and Silander Jr, J. A. 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. – Ecography 36: 1058–1069.
Merow, C., Smith, M. J., Edwards Jr, T. C., Guisan, A., McMahon, S. M., Normand, S., Thuiller, W., Wüest, R. O., Zimmermann, N. E. and Elith, J. 2014. What do we gain from simplicity versus complexity in species distribution models? – Ecography 37: 1267–1281.
Mertes, K. and Jetz, W. 2018. Disentangling scale dependencies in species environmental niches and distributions. – Ecography 41: 1604–1615.
Meyer, C., Kreft, H., Guralnick, R. and Jetz, W. 2015. Global priorities for an effective information basis of biodiversity distributions. – Nat. Commun. 6: 8221.
Mitchell, P. J., Monk, J. and Laurenson, L. 2017. Sensitivity of fine-scale species distribution models to locational uncertainty in occurrence data across multiple sample sizes. – Methods Ecol. Evol. 8: 12–21.
Moreno-Amat, E., Mateo, R. G., Nieto-Lugilde, D., Morueta-Holme, N., Svenning, J. C. and García-Amorena, I. 2015. Impact of model complexity on cross-temporal transferability in Maxent species distribution models: an assessment using paleobotanical data. – Ecol. Modell. 312: 308–317.
Moudrý, V. 2015. Modelling species distributions with simulated virtual species. – J. Biogeogr. 42: 1365–1366.
Moudrý, V. and Devillers, R. 2020. Quality and usability challenges of global marine biodiversity databases: an example for marine mammal data. – Ecol. Inform. 56: 101051.
Moudrý, V. and Šímová, P. 2012. Influence of positional accuracy, sample size and scale on modelling species distributions: a review. – Int. J. Geogr. Inf. Sci. 26: 2083–2095.
Moudrý, V., Komárek, J. and Šímová, P. 2017. Which breeding bird categories should we use in models of species distribution? – Ecol. Indic. 74: 526–529.
Moudrý, V., Keil, P., Gábor, L., Lecours, V., Zarzo-Arias, A., Barták, V., Malavasi, M., Rocchini, D., Torresani, M., Gdulová, K., Grattarola, F., Leroy, F., Marchetto, E., Thouverai, E., Prošek, J., Wild, J. and Šímová, P. 2023. Scale mismatches between predictor and response variables in species distribution modelling: a review of practices for appropriate grain selection. – Prog. Phys. Geogr. 47: 467–482.
Muscatello, A., Elith, J. and Kujala, H. 2021. How decisions about fitting species distribution models affect conservation outcomes. – Conserv. Biol. 35: 1309–1320.
Naimi, B., Skidmore, A. K., Groen, T. A. and Hamm, N. A. 2011. Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. – J. Biogeogr. 38: 1497–1509.
Naimi, B., Hamm, N. A. S., Groen, T. A., Skidmore, A. K. and Toxopeus, A. G. 2014. Where is positional uncertainty a problem for species distribution modelling? – Ecography 37: 191–203.
Newbold, T. 2010. Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. – Prog. Phys. Geogr. 34: 3–22.
Osborne, P. E. and Leitao, P. J. 2009. Effects of species and habitat positional errors on the performance and interpretation of species distribution models. – Divers. Distrib. 15: 671–681.
Papeş, M. and Gaubert, P. 2007. Modelling ecological niches from low numbers of occurrences: assessment of the conservation status of poorly known viverrids (Mammalia, Carnivora) across two continents. – Divers. Distrib. 13: 890–902.
Pearson, R. G., Raxworthy, C. J., Nakamura, M. and Townsend Peterson, A. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. – J. Biogeogr. 34: 102–117.

Peterson, A. T. 2014. Mapping disease transmission risk: enriching models using biogeography and ecology. – Johns Hopkins Univ. Press.

Peterson, A. T. and Samy, A. M. 2016. Geographic potential of disease caused by Ebola and Marburg viruses in Africa. – Acta Trop. 162: 114–124.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J. and Ferrier, S. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. – Ecol. Appl. 19: 181–197.

Proosdij, A. S. J. van, Sosef, M. S. M., Wieringa, J. J. and Raes, N. 2016. Minimum required number of specimen records to develop accurate species distribution models. – Ecography 39: 542–552.

Ramampiandra, E. C., Scheidegger, A., Wydler, J. and Schuwirth, N. 2023. A comparison of machine learning and statistical species distribution models: quantifying overfitting supports model interpretation. – Ecol. Modell. 481: 110353.

Ranc, N., Santini, L., Rondinini, C., Boitani, L., Poitevin, F., Angerbjörn, A. and Maiorano, L. 2017. Performance tradeoffs in target-group bias correction for species distribution models. – Ecography 40: 1076–1087.

Rattray, A., Ierodiaconou, D., Monk, J., Laurenson, L. J. B. and Kennedy, P. 2014. Quantification of spatial and thematic uncertainty in the application of underwater video for benthic habitat mapping. – Mar. Geod. 37: 315–336.

Raxworthy, C. J., Martinez-Meyer, E., Horning, N., Nussbaum, R. A., Schneider, G. E., Ortega-Huerta, M. A. and Townsend Peterson, A. 2003. Predicting distributions of known and unknown reptile species in Madagascar. – Nature 426: 837–841.

Reineking, B. and Schröder, B. S. 2006. Constrain to perform: regularization of habitat models. – Ecol. Modell. 193: 675–690.

Reside, A. E., Watson, I., VanDerWal, J. and Kutt, A. S. 2011. Incorporating low-resolution historic species location data decreases performance of distribution models. – Ecol. Modell. 222: 3444–3448.

Rhoden, C. M., Peterman, W. E. and Taylor, C. A. 2017. Maxent-directed field surveys identify new populations of narrowly endemic habitat specialists. – PeerJ 5: e3632.

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J. M., Jimenez-Valverde, A., Ricotta, C., Bacaro, G. and Chiarucci, A. 2011. Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. – Prog. Phys. Geogr. 35: 211–226.

Rocchini, D. et al. 2023. A quixotic view of spatial bias in modelling the distribution of species and their diversity. – NPJ Biodiv. 2: 10.

Sabatini, F. M. et al. 2021. sPlotOpen – an environmentally balanced, open-access, global dataset of vegetation plots. – Global Ecol. Biogeogr. 30: 1740–1764.

Santini, L., Benítez-López, A., Maiorano, L., Čengić, M. and Huijbregts, M. A. 2021. Assessing the reliability of species distribution projections in climate change research. – Divers. Distrib. 27: 1035–1050.

Segal, R. D., Massaro, M., Carlile, N. and Whitsed, R. 2021. Small-scale species distribution model identifies restricted breeding habitat for an endemic island bird. – Anim. Conserv. 24: 959–969.

Segurado, P. and Araujo, M. B. 2004. An evaluation of methods for modelling species distributions. – J. Biogeogr. 31: 1555–1568.

Seoane, J., Carrascal, L. M., Alonso, C. L. and Palomino, D. 2005. Species-specific traits associated to prediction errors in bird habitat suitability modelling. – Ecol. Modell. 185: 299–308.

Shiroyama, R., Wang, M. and Yoshimura, C. 2020. Effect of sample size on habitat suitability estimation using random forests: a case of bluegill, *Lepomis macrochirus*. – Ann. Limnol. Int. J. Limnol. 56: 13.

Sillero, N. 2011. What does ecological modelling model? A proposed classification of ecological niche models based on their underlying methods. – Ecol. Modell. 222: 1343–1346.

Sillero, N. and Barbosa, A. M. 2021. Common mistakes in ecological niche models. – Int. J. Geogr. Inf. Sci. 35: 213–226.

Sillero, N. and Gonçalves-Seco, L. 2014. Spatial structure analysis of a reptile community with airborne LiDAR data. – Int. J. Geogr. Inf. Sci. 28: 1709–1722.

Sillero, N., Arenas-Castro, S., Enriquez-Urzelai, U., Vale, C. G., Sousa-Guedes, D., Martínez-Freiría, F., Real, R. and Barbosa, A. M. 2021a. Want to model a species niche? A step-by-step guideline on correlative ecological niche modelling. – Ecol. Modell. 456: 109671.

Sillero, N., Dos Santos, R., Teodoro, A. C. and Carretero, M. A. 2021b. Ecological niche models improve home range estimations. – J. Zool. 313: 145–157.

Smith, A. B. and Santos, M. J. 2020. Testing the ability of species distribution models to infer variable importance. – Ecography 43: 1801–1813.

Smith, A. B., Murphy, S. J., Henderson, D. and Erickson, K. D. 2023. Including imprecisely georeferenced specimens improves accuracy of species distribution models and estimates of niche breadth. – Global Ecol. Biogeogr. 32: 342–355.

Soberón, J. and Nakamura, M. 2009. Niches and distributional areas: concepts, methods, and assumptions. – Proc. Natl Acad. Sci. USA 106: 19644–19650.

Støa, B., Halvorsen, R., Stokland, J. N. and Gusarov, V. I. 2019. How much is enough? Influence of number of presence observations on the performance of species distribution models. – Sommerfeltia 39: 1–28.

Stockwell, D. R. and Peterson, A. T. 2002. Effects of sample size on accuracy of species distribution models. – Ecol. Modell. 148: 1–13.

Stolar, J. and Nielsen, S. E. 2015. Accounting for spatially biased sampling effort in presence-only species distribution modelling. – Divers. Distrib. 21: 595–608.

Syfert, M. M., Smith, M. J. and Coomes, D. A. 2013. The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. – PLoS One 8: e55158.

Ten Caten, C. and Dallas, T. 2023. Thinning occurrence points does not improve species distribution model performance. – Ecosphere 14: e4703.

Tessarolo, G., Rangel, T. F., Araújo, M. B. and Hortal, J. 2014. Uncertainty associated with survey design in species distribution models. – Divers. Distrib. 20: 1258–1269.

Tessarolo, G., Ladle, R. J., Lobo, J. M., Rangel, T. F. and Hortal, J. 2021. Using maps of biogeographical ignorance to reveal the uncertainty in distributional data hidden in species distribution models. – Ecography 44: 1743–1755.

Thibaud, E., Petitpierre, B., Broennimann, O., Davison, A. C. and Guisan, A. 2014. Measuring the relative effect of factors affecting species distribution model predictions. – Methods Ecol. Evol. 5: 947–955.

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R. and Legendre, F. 2017. Taxonomic bias in biodiversity data and societal preferences. – Sci. Rep. 7: 9132.

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D. and Kadmon, R. 2007. A comparative evaluation of presence-only methods for modelling species distribution. – Divers. Distrib. 13: 397–405.

van Smeden, M., Moons, K. G., de Groot, J. A., Collins, G. S., Altman, D. G., Eijkemans, M. J. and Reitsma, J. B. 2019. Sample size for binary logistic prediction models: beyond events per variable criteria. – Stat. Methods Med. Res. 28: 2455–2474.

Varela, S., Anderson, R. P., García-Valdés, R. and Fernández-González, F. 2014. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. – Ecography 37: 1084–1091.

Velásquez-Tibatá, J., Graham, C. H. and Munch, S. B. 2016. Using measurement error models to account for georeferencing error in species distribution models. – Ecography 39: 305–316.

Veloz, S. D. 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. – J. Biogeogr. 36: 2290–2299.

Vollering, J., Schuiteman, A., de Vogel, E., van Vugt, R. and Raes, N. 2016. Phytogeography of New Guinean orchids: patterns of species richness and turnover. – J. Biogeogr. 43: 204–214.

Wang, L. and Jackson, D. A. 2023. Effects of sample size, data quality, and species response in environmental space on modeling species distributions. – Landscape Ecol. 38: 4009–4031.

Watcharamongkol, T., Christin, P. A. and Osborne, C. P. 2018. C4 photosynthesis evolved in warm climates but promoted migration to cooler ones. – Ecol. Lett. 21: 376–383.

Wieczorek, J., Guo, Q. and Hijmans, R. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. – Int. J. Geogr. Inf. Sci. 18: 745–767.

Williams, K. J., Belbin, L., Austin, M. P., Stein, J. L. and Ferrier, S. 2012. Which environmental variables should I use in my biodiversity model? – Int. J. Geogr. Inf. Sci. 26: 2009–2047.

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A. and NCEAS Predicting Species Distributions Working Group. 2008. Effects of sample size on the performance of species distribution models. – Divers. Distrib. 14: 763–773.

Wüest, R. O., Zimmermann, N. E., Zurell, D., Alexander, J. M., Fritz, S. A., Hof, C., Kreft, H., Normand, S., Cabral, J. S., Szekely, E., Thuiller, W., Wikelski, M. and Karger, D. N. 2020. Macroecology in the age of Big Data – where to go from here? – J. Biogeogr. 47: 1–12.

Xu, Q., Wang, X., Yi, J. and Wang, Y. 2024. Bias correction in species distribution models based on geographic and environmental characteristics. – Ecol. Inform. 81: 102604.

Zhang, G., Zhu, A. X., Huang, Z. P. and Xiao, W. 2018. A heuristic-based approach to mitigating positional errors in patrol data for species distribution modeling. – Trans. GIS. 22: 202–216.

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., Farooq, H., Herdean, A., Ariza, M., Scharn, R., Svantesson, S., Wengström, N., Zizka, V. and Antonelli, A. 2019. CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. – Methods Ecol. Evol. 10: 744–751.

Zizka, A., Antonelli, A. and Silvestro, D. 2021. Sampbias, a method for quantifying geographic sampling biases in species distribution data. – Ecography 44: 25–32.

Zurell, D. et al. 2020. A standard protocol for reporting species distribution models. – Ecography 43: 1261–1277.