npj | biodiversity

www.nature.com/npjbiodivers

PERSPECTIVE OPEN

Check for updates

A quixotic view of spatial bias in modelling the distribution of species and their diversity

Duccio Rocchini[1,2]✉, Enrico Tordoni[3], Elisa Marchetto[1], Matteo Marcantonio[4], A. Márcia Barbosa[5], Manuele Bazzichetto[2], Carl Beierkuhnlein[6], Elisa Castelnuovo[1], Roberto Cazzolla Gatti[1], Alessandro Chiarucci[1], Ludovico Chieffallo[1], Daniele Da Re[7], Michele Di Musciano[1,8], Giles M. Foody[9], Lukas Gabor[10,11], Carol X. Garzon-Lopez[12], Antoine Guisan[13,14], Tarek Hattab[15], Joaquin Hortal[16], William E. Kunin[17], Ferenc Jordán[18], Jonathan Lenoir[19], Silvia Mirri[20], Vítězslav Moudrý[2], Babak Naimi[21], Jakub Nowosad[22], Francesco Maria Sabatini[1,23], Andreas H. Schweiger[24], Petra Šímová[2], Geiziane Tessarolo[25], Piero Zannini[1] and Marco Malavasi[26]

Ecological processes are often spatially and temporally structured, potentially leading to autocorrelation either in environmental variables or species distribution data. Because of that, spatially-biased in-situ samples or predictors might affect the outcomes of ecological models used to infer the geographic distribution of species and diversity. There is a vast heterogeneity of methods and approaches to assess and measure spatial bias; this paper aims at addressing the spatial component of data-driven biases in species distribution modelling, and to propose potential solutions to explicitly test and account for them. Our major goal is not to propose methods to remove spatial bias from the modelling procedure, which would be impossible without proper knowledge of all the processes generating it, but rather to propose alternatives to explore and handle it. In particular, we propose and describe three main strategies that may provide a fair account of spatial bias, namely: (i) how to represent spatial bias; (ii) how to simulate null models based on virtual species for testing biogeographical and species distribution hypotheses; and (iii) how to make use of spatial bias - in particular related to sampling effort - as a leverage instead of a hindrance in species distribution modelling. We link these strategies with good practice in accounting for spatial bias in species distribution modelling.

npj Biodiversity (2023)2:10; https://doi.org/10.1038/s44185-023-00014-6

INTRODUCTION

> 'A greater acknowledgement of model uncertainty often has the consequence of widening our uncertainty bands [...]. Since hedging against uncertainty is hard work, this is an unpopular turn of events, at least in the short run. But [...] which is worse - widening the bands now, or missing the truth later?'[1]

Ecological processes are often spatially and temporally structured, so both environmental variables and species observations can potentially be autocorrelated[2,3]. Modelling the geographic distribution of species and the composition of ecological communities is key to preserve biodiversity and support a proper management of the habitats in which species live and have adapted over their evolutionary history[4–7]. From this point of view, predicting the distributions of species and communities in space and time provides a powerful tool for conservation planning[8–12]. Hence, studying species distribution changes might represent an effective approach to understand the complex interplay between the current biodiversity crisis and anthropogenic climate change[13–15].

Yet, complete knowledge of the distribution of any plant or animal species, and how these aggregate into more or less diverse communities, is hardly achievable. In some cases, the battle of ecologists against the many problems related to the modelling of species distributions becomes quixotic, or similar to fighting against a chimera. Hence, it needs to be approached in an idealistic way, fostering new ideas to fight the many challenges associated with biodiversity modelling[16]. Such a battle requires proper modelling approaches, which simultaneously account for empirical evidence[13] and stochastic processes[15]. In this context,

[1]BIOME Lab, Department of Biological, Geological and Environmental Sciences, Alma Mater Studiorum University of Bologna, via Irnerio 42, 40126 Bologna, Italy. [2]Czech University of Life Sciences Prague, Faculty of Environmental Sciences, Department of Spatial Sciences, Kamýcka 129, Praha - Suchdol 16500, Czech Republic. [3]Department of Botany, Institute of Ecology and Earth Science, University of Tartu, J. Liivi 2, 50409 Tartu, Estonia. [4]Evolutionary Ecology and Genetics Group, Earth and Life Institute, UCLouvain, 1348 Louvain-la-Neuve, Belgium. [5]CICGE (Centro de Investigação em Ciências Geo-Espaciais), Universidade do Porto, Porto, Portugal. [6]Biogeography, BayCEER, University of Bayreuth, Universitaetsstraße 30, 95440 Bayreuth, Germany. [7]Georges Lemaître Center for Earth and Climate Research, Earth and Life Institute, UCLouvain, Louvain-la-Neuve, Belgium. [8]Department of Life, Health and Environmental Sciences, University of L'Aquila, Piazzale Salvatore Tommasi 1, 67100 L'Aquila, Italy. [9]School of Geography, University of Nottingham, Nottingham, UK. [10]Dept of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA. [11]Center for Biodiversity and Global Change, Yale University, New Haven, CT, USA. [12]Knowledge Infrastructures, Campus Fryslan University of Groningen, Leeuwarden, The Netherlands. [13]Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland. [14]Institute of Earth Surface Dynamics, University of Lausanne, 1015 Lausanne, Switzerland. [15]MARBEC, Univ Montpellier, CNRS, Ifremer, IRD, Sète, France. [16]Department of Biogeography and Global Change, Museo Nacional de Ciencias Naturales (MNCN-CSIC), Madrid, Spain. [17]University of Leeds, Leeds, UK. [18]University of Parma, Parma, Italy. [19]UMR CNRS 7058 "Ecologie et Dynamique des Systèmes Anthropisés" (EDYSAN), Université de Picardie Jules Verne, 1 Rue des Louvels, 80000 Amiens, France. [20]Department of Computer Science and Engineering, Alma Mater Studiorum University of Bologna, via Irnerio 42, 40126 Bologna, Italy. [21]Rui Nabeiro Biodiversity Chair, MED Institute, University of Évora, Évora, Portugal. [22]Institute of Geoecology and Geoinformation, Adam Mickiewicz University, Krygowskiego 10, 61-680 Poznan, Poland. [23]Faculty of Forestry and Wood Sciences, Czech University of Life Sciences Prague, Prague - Suchdol, Czech Republic. [24]Department of Plant Ecology, Institute of Landscape and Plant Ecology, University of Hohenheim, Stuttgart, Germany. [25]Federal University of Goiás, Campus Central, Anápolis, Brazil. [26]University of Sassari, Department of Chemistry, Physics, Mathematics and Natural Sciences, Sassari, Italy. ✉email: duccio.rocchini@unibo.it

Species Distribution Models (hereafter SDMs, also known as Ecological Niche Models, Habitat Suitability Models and many other names used in the scientific literature[9,17]) are powerful tools, since they provide insights into species or community distributions in space and their potential shifts over time[18]. In practical terms, depending on the final interest or overarching goal, the label might change, e.g., labelling it SDMs when the focus is on the spatial distribution of species and labelling it ENMs when the focus is on the underlying drivers, namely the niche requirements of species[19].

In this paper we did not explicitly distinguish between SDMs, ENMs or HSMs, the three main acronyms used for the same underlying model machinery, since they are all relying on the estimation of ecological requirements of species for predicting their distributions in space and time[19]. In addition to that, other labels such as Potential Habitat Distribution Models (PHDMs), Climate Envelope Models (CEM), Resource Selection Functions (RSF) and others are also used to name this category of niche- or habitat-suitability based distribution models. In fact, we share the view that niches—or habitats—should not be distinguished from or opposed to distributions because these two are faces of the same coin, where the coin is a species with one side being the distribution within the geographical space and the other side the niche as an envelope of habitat suitability within the environmental space. Hence, in our opinion, niche, habitat suitability and distribution are too much entangled to dissociate them into separate entities or types of models (see ref. [20] for an example of such an entanglement focusing only on SDMs and ENMs).

All SDMs typically rely on (i) species distribution data, either in the form of both presence and background data (also called pseudo-absences) or presence and true absence data gathered in the field, as well as (ii) a list of predictor variables expected to represent the ecological and geographical drivers of the species' distribution range[17]. Understanding the spatial covariation of species and their assembly into communities is crucial in ecology[21,22], so a wealth of methodological approaches has been developed recently to account for species co-occurrences; for instance, joint SDMs—for modelling the covariance of multiple species together (e.g., ref. [23])—or stacked SDMs—for modelling single species distributions sequentially and combining them afterwards (e.g., ref. [24])—can be used to estimate community-level parameters like species richness[25,26]. In other words, properly stacking SDMs and considering the biotic interactions among species[24,27] will yield more realistic estimates of spatial patterns in alpha diversity that relate to environmental gradients[28]. This said, no modelling techniques are free from the uncertainty coming from biases in the input data, like uneven sampling effort[29–35] or spatial positioning errors[36–38]. Here, an integration of species distributions and community-level biodiversity modelling can be performed under the Spatially-Explicit Species Assemblage Modelling framework (SESAM[39]), in which species associations and biotic interactions are explicitly considered[40].

The increasing availability of spatially explicit open-access databases on species distribution and community composition (e.g., GBIF, sPlotOpen) with appropriate georeferencing[41–43]—coupled with technical and methodological advances for data querying, cleaning, and analysis—opened up new opportunities for global species distribution modelling[44–46]. Furthermore, large-scale environmental layers describing bioclimatic and edaphic conditions have been effectively used as proxies of ecological and climatic drivers of species distributions[47]. These global gridded data, under a structured framework, have been used to systematically select proper environmental variables from a large suite of spatio-environmental variables[48]. Nonetheless, the actual knowledge on species distributions over wide geographical regions is still far from being complete[49–54], and suffers from pervasive geographical biases[55–64].

Projecting species distributions for regions and time periods other than those used during model calibration (i.e., model extrapolation)—based on, e.g., bioclimatic variables—requires explicit recognition of all the possible sources of spatial bias, or the use of mechanistic models of species distribution[65]. In fact, transferring model rules onto non-analogous bioclimatic conditions is perilous and a very risky business[66–68]. In other words, extending such projections to new regions involves some sort of extrapolation risk, simply because the recorded occurrences used for model calibration are incomplete or spatially biased, thus increasing spatial uncertainty[69–71]. For instance, methods would be needed to minimize the effects of spatial autocorrelation among records within the geographical space[72–75], although in some cases spatial autocorrelation could have minimal effects in peculiar regions, such as in topographically rugged landscapes[76,77]. More generally, starting from spatially biased in-situ samples (or predictors), undesired model outcomes can be expected[78].

This paper aims at addressing the spatial component of data-driven biases in species distribution modelling, and at proposing potential solutions to explicitly test and account for it. Our major goal is not to propose existing or new methods to remove spatial bias from the modelling procedure, which would be impossible without a proper knowledge of all the processes generating it, but rather to propose alternatives to explore and handle it. In particular, we describe three main strategies that may provide a fair account of spatial bias, namely: (i) how to represent spatial bias; (ii) how to build null models based on virtual species for testing biogeographical and species distribution hypotheses; and (iii) how to make use of spatial bias—in particular related to sampling effort—as a leverage instead of a hindrance in species distribution modelling. In each one of these sections we outline what would be good practices to account for spatial bias in species distribution modelling.

VISUALIZING SPATIAL BIAS IN THE DISTRIBUTION OF SPECIES AND THEIR DIVERSITY

Recently, the massive increase in the availability of biodiversity data[43], coupled with enhanced computing power and modelling techniques, has fostered a new wave of large-scale analyses of biodiversity patterns[45,79,80]. Nonetheless, data quality plays a crucial role in this process[54,81]. In fact, biodiversity knowledge is often skewed toward specific taxonomic groups[82], wealthy regions of colonial history[64], English-speaking research[83,84], and/or environmental domains[14,82,85,86], which are the major issues among the so-called seven shortfalls of biodiversity data[59].

The undersampling of some geographical areas—named the 'Wallacean shortfall' by Lomolino[49] (see also ref. [87])—was recently recognized as one of the main factors preventing an exhaustive large-scale understanding of biodiversity patterns[54,88]. Even when biodiversity data are available for a well-studied taxonomic group, these might suffer from a number of bias sources, just to cite a few[33,75,89–91]: lack of standardized sampling design, inconsistent spatial scales, inadequate environmental coverage of the surveys, and observer's/recorder's bias (e.g., proximity to roads). Indeed, the large variety of standardized and unstandardized sampling schemes used to survey the distribution of different biological groups often adds up as an additional source of heterogeneity in the data, which may increase the spatial bias and thus affect the complex exercise of modelling species distributions. Likewise, site accessibility and proximity to roads, also have strong effects on data quality, biodiversity inventories being more intensive in locations closer to research centres, infrastructure, highways or places allowing easier access[92–96]. Moreover, the striking geographic bias in the accessibility to resources and in scientific data processing among different regions across the globe can only increase gaps in the data. Altogether, bias in data quality

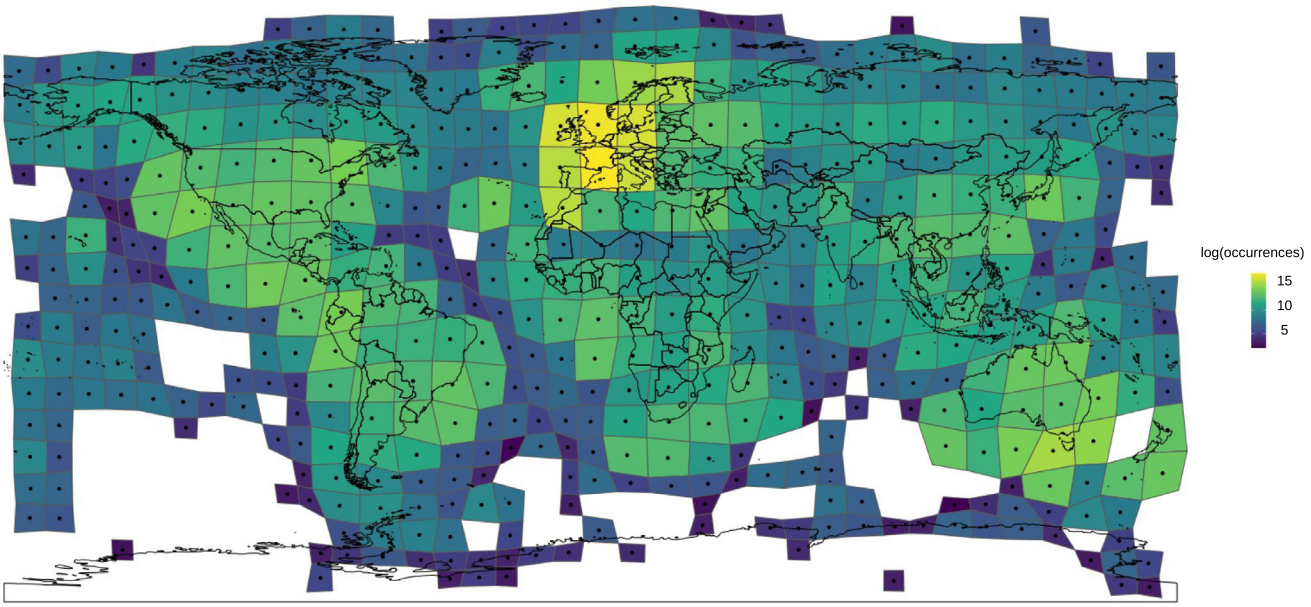

Fig. 1 Plant species occurrences over the globe available in GBIF (https://www.gbif.org, latest access: December 2021). The cartogram or density-equalizing map as proposed by Dorling[200] and Gastner and Newman[201]) shows a bias on species occurrences towards continents with higher sampling effort. To generate the cartogram, a geographical grid of 10 degrees was superimposed on the dataset and the grid cells were further distorted according to the amount of plant species occurrences.

represents a key issue in current macroecological and biogeographical research[97], and hinders realizing the full potential of using large-scale databases in biodiversity modelling (ref. [59,96,98,99], see Fig. 1). Examples exist where smooth geographical biases could be controlled during modelling procedures, in case some in-situ data have still been sampled even in remote areas[100], but spatial lack of information and stronger bias is generally expected to severely hinder final results[101].

Spatial bias has been shown to increase uncertainty in the data[91], and has strong effects on the outcomes of the whole modelling process[78,102,103]. For example, species-people correlations, in which more populated regions show higher biodiversity simply because they are more thoroughly surveyed, are now well known[104]. Ensuring an adequate sampling design is of utmost importance to avoid the generation of truncated species response curves[31,105]. For instance, datasets biased towards widespread environmental conditions across the study area[106–108] hampered the characterization of species responses to the effects of land transformation or the rarest climate conditions in highly dynamic landscapes such as the Brazilian Atlantic Forest[109]. Moreover, the autecology of species and related eco-geographic characteristics such as species traits[110], range size[111] and species niche breadths (i.e., generalist vs specialist[112]), among other factors, can ultimately influence the performance of species-related models[35].

Recent methodological advances have been proposed to limit spatial bias in data distribution. These can apply either when sampling new species data, or by resampling available data inside a strongly biased dataset[30,43,113], and restricting analyses to the geographical regions holding enough data coverage[114]. For instance, Hattab et al.[115] developed a scheme that, by ensuring a systematic sampling of field observations within the environmental conditions available across the study area, can aid in limiting potential shortcomings when modelling species distribution while being not in equilibrium with the contemporary environment (e.g., the case of a recent introduction of an invasive alien species). Likewise, Lembrechts et al.[116] developed a new framework to design standardized microclimate networks able to capture the largest variation in microclimate at regional or national extents[117,118].

To appropriately map large-scale patterns of species distributions, the spatial structure of sampling bias must be first understood[63]. For instance, direct gradient analysis[119] might be used to relate the sampling effort of a focal species distribution with the assumed continuous variation of spatial predictors[120]. In some cases, spatial bias can be attenuated by (i) reducing the clustering of presences within the geographical space[108] using approaches such as spatial data thinning[121] or background thickening[34], or (ii) tuning the model before predicting species distributions[122]. For instance, even in the case of data which are geographically biased, regularization of the models can lead to high quality outputs. As an example, when clumping depends on sampling bias, using spatial or environmental filtering[123] or rarefaction methods before running SDMs may amend the final output[124]. Concerning spatial data thinning[33], it might decrease the probability of retaining species with unique environmental conditions. However, in case of a gradual species response to environmental gradients, there is a high model sensitivity to an inappropriate use of data thinning in the environmental space, based on e.g., thresholding methods[125]. From this point of view, a blind data thinning without testing model sensitivity is strongly discouraged. Hence, for instance, proper model averaging might reduce prediction errors[126]. Besides, the combination of predictions derived from different algorithms has generated much attention under the ensemble models umbrella[127], although in some cases ensemble models might not outperform well-tuned individual models based on machine learning algorithms such as Random Forests or Boosted Regression Trees[128].

Another important effect of sampling bias is that it creates information gaps[129,130]. This could be solved with recourse to citizen science, although it is well known that such information is even more biased (i) spatially, e.g., with a higher amount of data near roads, cities, research centres, in peculiar ecosystems or regions and, more globally, in the northern hemisphere, but also (ii) taxonomically, toward certain charismatic groups, e.g., vertebrates in terrestrial ecosystems[94,96,131]. In order to solve sampling completeness issues, new tools are now available based on diversity estimates and further fine-tuning of datasets, before they are used for further analysis. As an example, Lobo et al.[132]

propose a tool to estimate the degree of completeness in biodiversity surveys in each territorial unit, when the number of records (including repeated species) is available, as a surrogate of sampling effort. After having estimated the relationship between the number of records and cumulative species richness, Lobo et al.[132] suggest that the slope of the species accumulation curves and completeness percentages can be used to distinguish and map the level of survey per territorial unit. A similar approach has been proposed by Mokany et al.[133] based on alpha- and beta-diversity models to measure data completeness. When the number of records is not available for each territorial unit, another approach consists in dividing the study area into regions with known differences in the levels of survey effort. Models can then be computed on these different regions, to check if the observed relationships are consistent among them[104], obviously provided that all the considered regions span the entire species niche to avoid niche truncation[105,134].

Finally, it is also of primary importance to reveal the uncertainty in distributional data underlying SDMs, which can be achieved by maps of ignorance accounting for different sources of errors, such as data quality, time elapsed among the field observations, inventory completeness and the eco-geographic distance between species presences and absences (including true absences or pseudo-absences)[53,75,101,135,136]. More recently, Konig et al.[137] suggested a framework to increase the integration of biodiversity data across domains and resolutions (e.g., from point occurrences to entire floras) for scalable and integrative biodiversity research, especially when the quality of primary data can be integrated with expert knowledge[138].

USING VIRTUAL SPECIES TO HIGHLIGHT POTENTIAL SPATIAL BIASES OF SDMS

In most cases, there is no complete information about the 'reality' of the focal species distribution besides the data collected in-situ[101]. This is partly because the completeness of the data extracted from surveys (recorded in-situ) is difficult to measure[139].

For instance, occurrence data from natural history collections, such as museum or herbaria collections, tend to be very incomplete with a relatively high amount of false absences—i.e., species occurrences missed by the observer in the field in case of a rare or difficult to identify species (see ref. [140] on detection bias). Such incompleteness affects our ability of detecting the real spatial coverage of the samples and records available for modelling[141]. These limitations, in turn, can seriously flaw final results of species distribution models, by distorting the relationship between species occurrences and the underlying environmental patterns[56,142]. Yet, quantifying sources of error is essential for proper descriptive or mechanistic modelling of species distributions[143].

Making use of simulated or in-silico datasets—the so-called 'virtual ecologist' approach[143]—allows to generate distribution data with known ecological characteristics[76], considering that virtual species are better at rejecting candidate models than they are at supporting them[143–147]. The use of virtual species is burgeoning in ecology to build toolkits implementing in-silico analytical experiments simulating natural processes, thanks to the complete control on the configurations of factors constraining the distributions of species[19]. Moreover, virtual species allow creating simulated data for benchmarking models of different complexity. This is true passing from traditional SDMs projecting simple distributions, to those including population dynamics (the so-called hybrid models[148], see also ref. [149] on population dynamics and regulation), up to hierarchical Bayesian process-based dynamic range models[150], considering that model complexity can impact the projection of species distributions[151].

Making use of virtual species data allows (i) controlling for random variation in species distributions as well as (ii) simulating patterns of distribution based on known relationships with, for instance, climatic variables (i.e., by species response curves). Due to the artificial nature of such data, the expected underlying processes shaping species distribution patterns can be adjusted or, at least, balanced to account for random or systematic noise[152]. The use of such spatially explicit simulated data helps reaching a better conceptualization and implementation of modelling techniques, leading to the creation of a dominant paradigm for robust generalization and further recommendations for conservation planning. This is difficult with empirical studies, mainly due to confounding effects of interactions among different data types, environmental variables, and methodologies to assess model accuracy[145,153]. Further, models simulating virtual scenarios based on different ecological processes can be used to assess the sensitivity of different SDM algorithms to the effects of historical processes on species distributions[154].

From this point of view, open-source spatial algorithms have been developed and are freely downloadable (e.g., refs. [152,155,156]). We also provide an example in R in Figs. 2 and 3, with the complete code in Appendix 1 or in the following GitHub repository: https://github.com/duccioroccchini/Virtual_species_SDM/(see also ref. [20] for a similar example). The concept of virtual species is not the only example of virtual individuals/surfaces, since it has been widely used in disciplines other than ecology—e.g., in geology, virtual globes have been used for geophysical modelling[157].

Passing from species to assemblages, virtual communities can be simulated (Figs. 2 and 3) to understand what should be an effective sampling effort to predict the distribution of species assemblage, for instance when stacking separate species distribution models[146]. This is generally done by simulating virtual species in a community given a certain virtual species richness, and then manipulating this artificial set by changing different sampling parameters such as sample size, sampling strategy or different species distribution modelling algorithms such as Generalized Linear Models, Generalized Additive Models, MaxEnt, Boosted Regression Trees or Random Forests[146]. This approach is particularly useful, since it allows to better understand species co-existence, which is a long-lasting theme[158] and (still) an open question[159,160] in ecology. Furthermore, simulations of different sampling design strategies by virtual communities represent a solid basis for developing experimental designs, which guarantee a high reproducibility and avoid low statistical power due to e.g., small sample size[152].

Operationally speaking, hitherto there is no consensus about the best methods for generating virtual species distributions. Various examples exist based on: (i) model-based simulations; (ii) model fitting to in-situ data; or (iii) predefined theoretical response (see ref. [76]). In some cases, it is possible to combine several virtual species to compose a community[146]. Starting from a set of environmental combinations, e.g., using a Principal Components Analysis (PCA) to reduce the number of dimensions of the environmental space, the overlap among niches of different virtual species can be set and controlled to look at potential complements with a focal species of interest[161]. This procedure allows understanding patterns at the community level and balancing potential spatial sampling bias related to rare species. A complete review on the backbone of the virtual species approach is provided by Miller[144] and Meynard et al.[147]. An experimental approach to data science requires that simulations are a key elements of experimental tests[162–164]. In this paper, we provided an operational way of generating virtual species; albeit we rely on a synthetic and simplistic community of four virtual species, more complex communities composed by thousands of virtual species can be created[165–168]. Further, there is already a broad spectrum of methods for implementing virtual species[147].

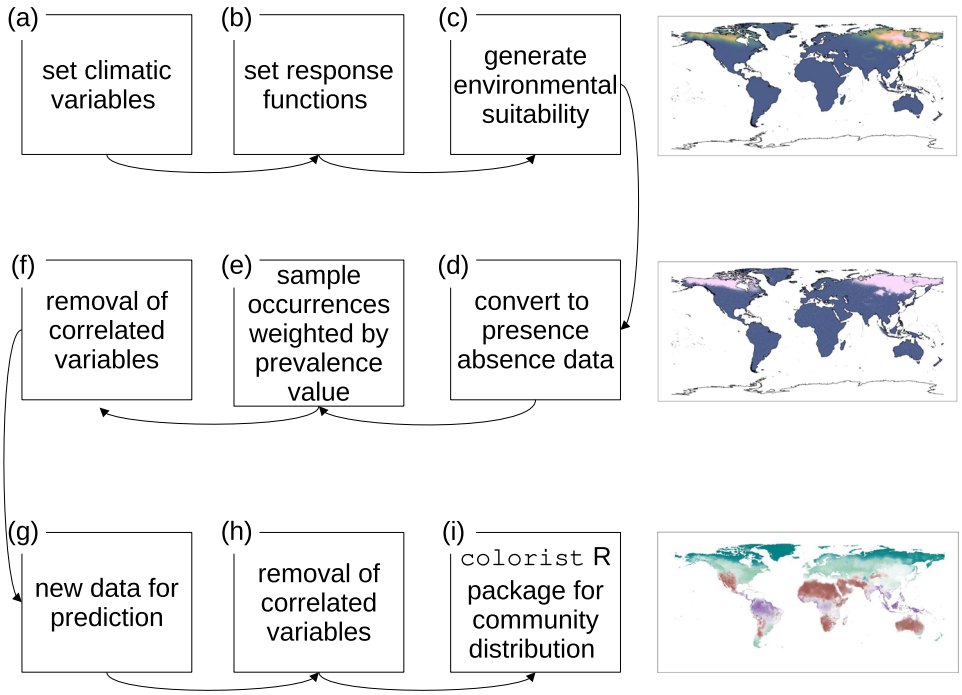

Fig. 2 The procedure used to generate virtual species and colorist-based community distribution. First of all, the climatic variables are selected (**a**) and the species response functions of each environmental variable are set (**b**). The environmental suitability of the virtual species distribution is generated in conformity with the response functions (**c**). Then, a logistic conversion transforms it into presences and absences (**d**) and presence and absence points are sampled according to the sample prevalence value (**e**). Furthermore, a collinearity test is performed and the correlated variables are removed (**f**). Once the statistical model has been calibrated, the climatic variables for the prediction are selected (**g**) and—among them—those which are correlated are deleted (**h**). Eventually, multiple virtual species distributions are combined together in `colorist` R package to map community distribution (**i**). Results are shown in Fig. 3. The complete code to generate virtual species and final maps is available in both Appendix 1 and at the following GitHub repository link: https://github.com/ducciorocchini/Virtual_species_SDM/.

SAMPLING EFFORT BIAS AS A COVARIATE IN SDMS

Uneven sampling effort is a crucial source of spatial bias. For instance, many areas over the planet are oversampled due to their higher accessibility and closeness to research institutes and universities. On the other hand, most remote areas are undersampled, mainly due to inaccessibility and/or inhospitality to humans[92,169]. The effect of this spatial non-stationarity (see refs. [170–172]) is a spatial bias in the perceived species distribution and diversity patterns over the planet[173–175], and therefore a limited coverage of niche-based responses to the environment for many species[56]. If undersampled areas are included in the modelled region, such spatial bias can lead to zero inflation—related to true or false absences in the data—which is problematic to handle[176]. Flexible methods are therefore required to face data with proportions of zeros larger than those expected from pure count Poisson data[177]. This said, zero inflation is not necessarily due to a bias in the species data, but it is often simply an inherent property of ecological systems, where a large number of species are infrequent or rare. Individuals belonging to rare and/or elusive species might be missed, also depending on the strategy of the sampling design adopted. In other words, species distribution models are expected to show a diverse sensitivity to sampling effort, depending on the taxonomic group whose distribution they attempt to forecast[178].

Unbiased estimates of species distributions are strictly related to the assumption of a random distribution of sampling effort over the area under study. This is also true considering that, when using SDMs to make inference, any model is wrong in its intrinsic definition[176], but some are less wrong than others and can still provide useful outputs. Sampling effort is also inherently related to scale: species occurrence and community diversity are generally scale dependent. Various approaches have been used to investigate the scale-dependency of ecological variables, from nested sampling[179] to distance-based sampling[176]. However, these do not guarantee that sampling effort is explicitly measured and/or controlled for. This is particularly true when considering the covariance of different variables[180]—in our case, as an example, of different species. Using mixed-effects or hierarchical models in SDMs, e.g. grounded in the spatial Mixed-effects Models (spaMM) framework[181,182], should help solving such bias by accounting for pseudo-replication issues.

Obviously, additional causes of uncertainty might increase the spatial bias of species distribution models. For instance, taxonomic misidentification and phenological mismatches of species can lead to highly unreliable models if the biological subject of analysis and the sampling period are not adequately defined[183], e.g., by sampling a site at the wrong time period or by using an outdated taxonomy[184]. Yet, while these and other sources of uncertainty have non-negligible effects on SDMs accuracy, their impact is normally smaller compared to that of sampling effort[185], as it may mainly affect the interpretation of the resulting models[183].

Accounting for uncertainty in SDMs may increase their reliability and predictive power[186]. Based on the above, making use of sampling effort estimates as covariates directly into SDMs can certainly increase their accuracy[174,187–189]. These estimates of sampling effort can be based on (i) the accessibility of the surveyed areas; (ii) time spent on single plots; (iii) multiple visiting periods to catch the right phenological period; (iv) the number of records (including repeated species) per territorial unit; or (v) the number of occurrences within the same taxonomic group, e.g. the genus or family that the focal species belongs to. Such estimates

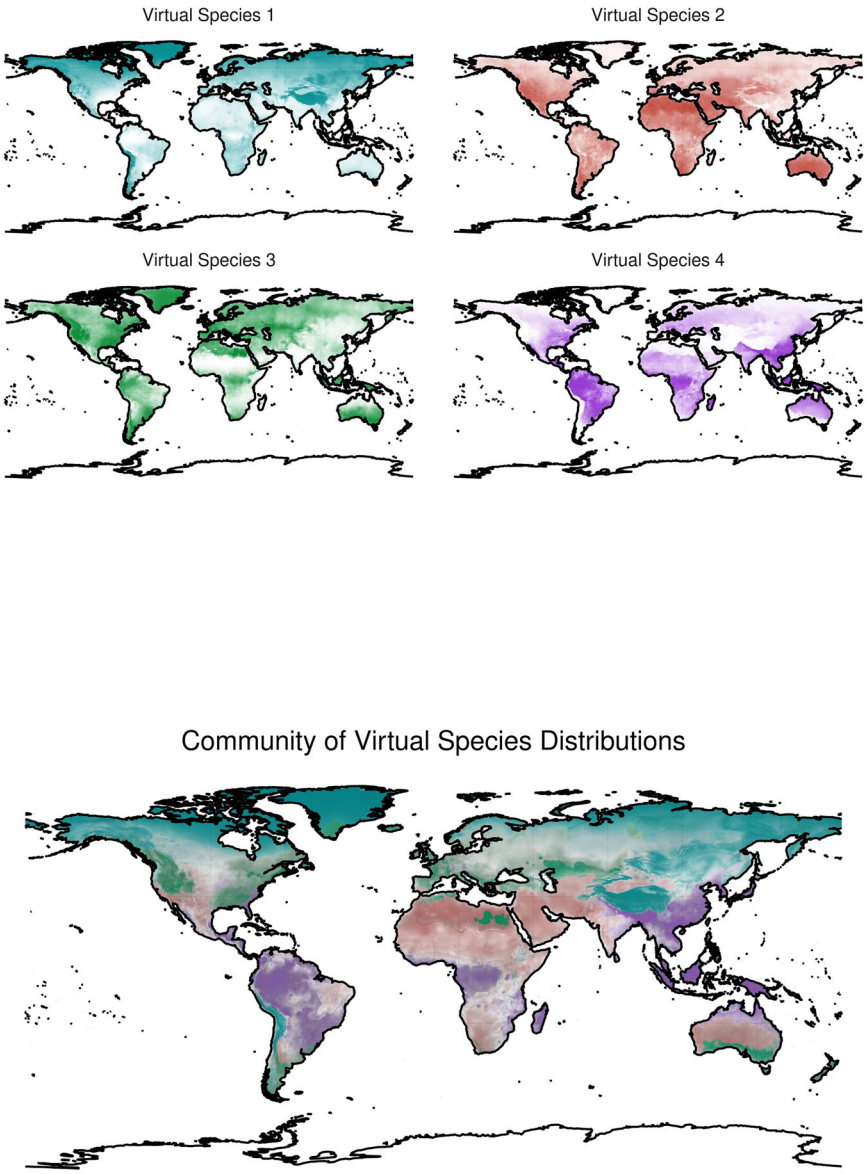

Fig. 3 Virtual species can be built to form a virtual community. Starting from colours of single virtual species distributions and relying on the `colorist` package, it is possible to spatially merge colours and their overlaps in a final gamut which account for single species colour intensity.

of sampling effort can then be included as covariates in the analysis[190–192]. Similarly, estimates of completeness (e.g., ref. [132]) or multivariate estimates of data-driven uncertainty, such as the previously cited maps of ignorance approach[75,101], can be used as ancillary predictors in SDMs, or as spatially-explicit error terms in regression-based modelling techniques[186].

Starting from the intuitive assumption that a higher sampling effort could be related to intrinsically higher prevalence of species' occurrence data inside a region, Bayesian inference can integrate this information in the modelling of species distributions to guide model predictions. However, Bayesian methods are, in general, computationally intensive, which makes them sometimes unfeasible for many species over large areas. Alternatively, one can generate very simple covariates to capture the effect of sampling effort in traditional SDMs. For instance, Wasof et al.[193] fitted SDMs for vascular plant species that included several covariates: a region effect (Alps vs. Fennoscandia) to test potential differences in distribution patterns between the two investigated regions and a covariate reflecting sampling effort based on the total number of presence/absence records available per sampled grid cell (1 km^2) to account for the spatially imbalanced data within each of the two investigated regions. Furthermore, Rocchini et al.[175] included sampling effort as a hyper-prior in a multilevel model structure, by considering different degrees of association between sampling effort at large spatial extents to predict the probability of species presence (*Abies alba* over Europe) in smaller nested areas. Sampling effort was estimated as the number of revisiting dates and used for further modelling in three main manners: no effect, mild effect and strong effect. The model with the strongest importance assigned to sampling effort significantly corrected final results for sampling effort bias (Fig. 4). This indicates that sampling effort might be used to supplement the often incomplete information provided by species presence at fine spatial scales. This modelling approach could also be extended considering similar species characterized by opposite degrees of sampling effort in an area (or even the overall species sampling effort; see ref. [194]). Data on sampling effort for a well surveyed and widespread species could also be considered to correct model

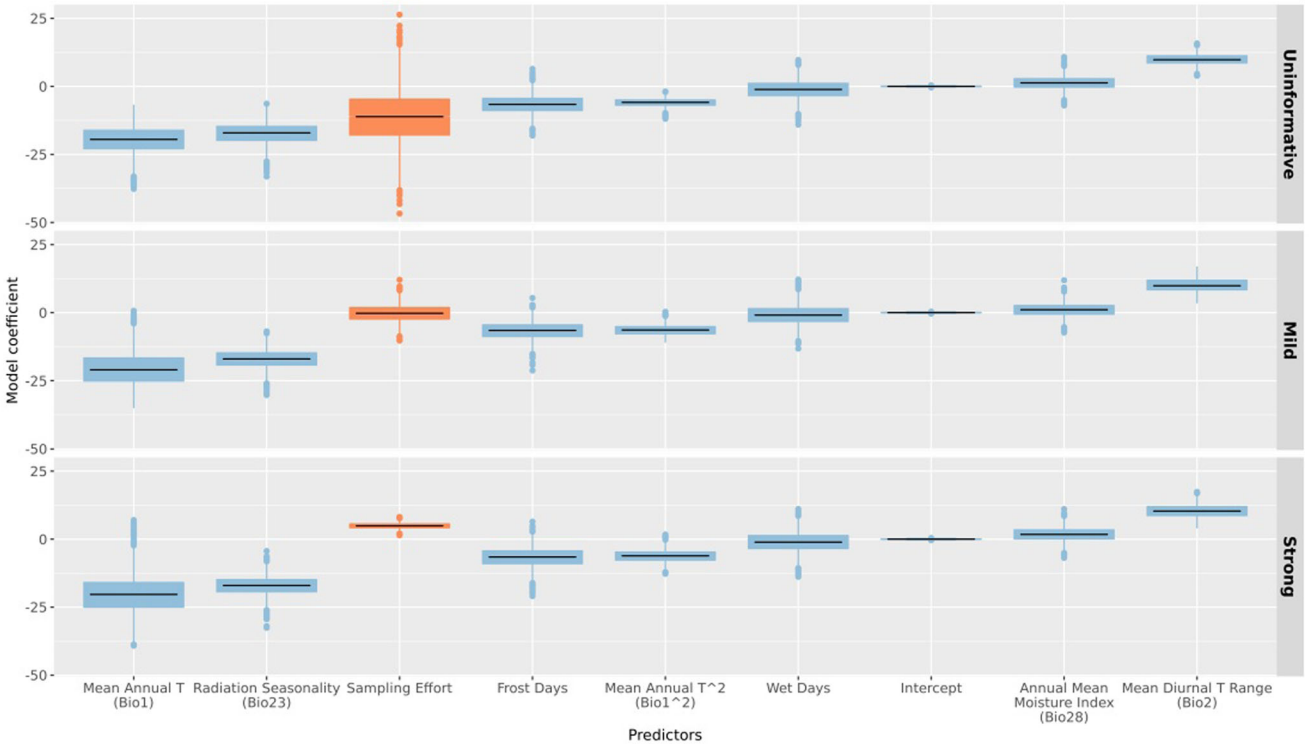

Fig. 4 Boxplots of the β coefficients in three different models using a different prior on sampling effort. Each box represents the 1st and 3rd quartiles of a coefficient distribution, the black horizontal line the distribution median, the whiskers the limits of the 1.5*interquartile range, while the filled circles represent the outlying points. We showed in red the boxplots reporting the distribution of the β coefficient of the sampling effort. Relying on Bayesian statistics it is possible to set three priors on sampling effort: not considering its effect, considering its effect in a mild manner, or in a strong manner. Sampling effort can be measured as an example by the number of revisiting dates. The precision of sampling effort increased passing from the model with an uninformative prior on sampling effort, through that with a mild prior, reaching its highest value in the model with a strong prior. Controlling sampling effort bias using a strongest prior could lead to the comparability of models related to species with opposite degrees of sampling effort in an area. See the main text for additional information. From Rocchini et al.[175]: License Number: 5495740269939, License date: Feb 25th 2023, Licensed Content Publisher: Elsevier.

outputs for a similar, but less sampled, species belonging to, e.g., the same genus[102].

CONCLUSION

In this short essay, we have addressed a range of methods to quantify and account for spatial bias when mapping species distribution and diversity (see also ref. [195]). Based on this general overview of the issues related to spatial bias in modelling species distribution, we basically propose (i) to integrate several methods to set the best tuning and achieve optimal model complexity when modelling distributions of species and their relative diversity[196,197] as well as (ii) to find the most effective visualization techniques to explore model behaviour[198].

If left unchecked, spatial bias could impair species distribution models/outputs, thereby resulting in pervasive biases along SDMs of different species, as spatially-structured sampling biases are often shared by all species pertaining to the same group. Implementing robust methods to map species distributions and spatial bias is crucial for natural resource management. In particular, two critical points must be faced explicitly: (i) integrating prior knowledge for improving the prediction of species distributions over wide geographical areas, and (ii) quantifying and visualizing the uncertainty associated with species distribution predictions over large geographical scales. Improved knowledge in areas where the modelled species are predicted to spread, along with illustration of uncertainty of predictions in an easily interpretable map, can lead to more effective management strategies[199]. This would allow timely actions to be initiated, both in case of the protection of natural species and the management of invasive species.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The virtual data used in this paper is free and open and can be generated by the code provided in section "Code availability". The full set of empirical data can be downloaded at https://www.gbif.org/.

CODE AVAILABILITY

The code used in this paper is available in Appendix 1 and at the following GitHub repository: https://github.com/ducciorocchini/Virtual_species_SDM/.

Received: 8 June 2022; Accepted: 23 March 2023;
Published online: 03 May 2023

REFERENCES

1. Draper, D. Assessment and propagation of model uncertainty. *J. R. Stat. Soc. Ser. B* **57**, 45–97 (1995).
2. Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J. & Bretagnolle, V. Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Glob. Ecol. Biogeogr.* **23**, 811–820 (2014).

3. Pereira, J., Saura, S. & Jordan, F. Single-node vs. multi-node centrality in landscape graph analysis: key habitat patches and their protection for 20 bird species in NE Spain. *Methods Ecol. Evol.* **8**, 1458–1467 (2017).
4. Van Horne, B. Density as a misleading indicator of habitat quality. *J. Wildlife Manag.* **47**, 893 (1983).
5. Ricotta, C., Godefroid, S. & Rocchini, D. Patterns of native and exotic species richness in the urban flora of Brussels: rejecting the "rich get richer" model. *Biol. Invasions* **12**, 233–240 (2010).
6. Marcantonio, M., Rocchini, D., Geri, F., Bacaro, G. & Amici, V. Biodiversity, roads, & landscape fragmentation: Two Mediterranean cases. *Appl. Geogr.* **42**, 63–72 (2013).
7. Newmark, W. D., Jenkins, C. N., Pimm, S. L., McNeally, P. B. & Halley, J. M. Targeted habitat restoration can reduce extinction rates in fragmented forests. *Proc. Natl Acad. Sci.* **114**, 9635–9640 (2017).
8. Guisan, A. & Thuiller, W. Predicting species distribution: offering more than simple habitat models. *Ecol. Lett.* **8**, 993–1009 (2005).
9. Guisan, A. et al. Predicting species distributions for conservation decisions. *Ecol. Lett.* **16**, 1424–1435 (2013).
10. Lecours, V., Gabor, L., Edinger, E. and Devillers, R. Fine-scale habitat characterization of The Gully, the Flemish Cap, and the Orphan Knoll, Northwest Atlantic, with a focus on cold-water corals. In *Seafloor Geomorphology as Benthic Habitat* (eds. Harris, P., Baker, E) 735–751 (Elsevier, 2020).
11. Santini, L., Benitez-Lopez, A., Maiorano, L., Cengic, M. & Huijbregts, M. A. Assessing the reliability of species distribution projections in climate change research. *Divers. Distrib.* **27**, 1035–1050 (2021).
12. Segal, R. D., Massaro, M., Carlile, N. & Whitsed, R. Small-scale species distribution model identifies restricted breeding habitat for an endemic island bird. *Anim. Conserv.* **24**, 959–969 (2021).
13. Dallas, T. A. & Hastings, A. Habitat suitability estimated by niche models is largely unrelated to species abundance. *Glob. Ecol. Biogeogr.* **27**, 1448–1456 (2018).
14. Lenoir, J. et al. Species better track climate warming in the oceans than on land. *Nat. Ecol. Evol.* **4**, 1044–1059 (2020).
15. Bokma, F., Bokma, J. & Monkkonen, M. Random processes and geographic species richness patterns: Why so few species in the north? *Ecography* **24**, 43–49 (2001).
16. Schwartz, M. A. The importance of stupidity in scientific research. *J. Cell Sci.* **121**, 1771–1771 (2008).
17. Guisan, A., Thuiller, W. & Zimmermann, N.E. Habitat Suitability and Distribution Models: With Applications in R. (Cambridge University Press, 2017).
18. Bittner, T., Jaeschke, A., Reineking, B. & Beierkuhnlein, C. Comparing modelling approaches at two levels of biological organisation - Climate change impacts on selected Natura 2000 habitats. *J. Veg. Sci.* **22**, 699–710 (2011).
19. Saupe, E. E. et al. Variation in niche and distribution model performance: The need for a priori assessment of key causal factors. *Ecol. Modell* **237–238**, 11–22 (2012).
20. Inman, R., Franklin, J., Esque, T. & Nussear, K. Comparing sample bias correction methods for species distribution modeling using virtual species. *Ecosphere* **12**, e03422 (2021).
21. Thompson, J. N. Variation in interspecific interactions. *Annu. Rev. Ecol. Syst.* **19**, 65–87 (1988).
22. Pereira, J., Battiston, F. & Jordan, F. Priority areas for protection of plant-pollinator interaction networks in the Atlantic Forest. *Ecol. Indic.* **136**, 108598 (2022).
23. Tobler, M. W. et al. Joint species distribution models with species correlations and imperfect detection. *Ecology* **100**, e02754 (2019).
24. Gavish, Y. et al. Accounting for biotic interactions through alpha-diversity constraints in stacked species distribution models. *Methods Ecol. Evol.* **8**, 1092–1102 (2017).
25. Norberg, A. et al. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecol. Monogr.* **89**, e01370 (2019).
26. Zurell, D. et al. Testing species assemblage predictions from stacked and joint species distribution models. *J. Biogeogr.* **47**, 101–113 (2020).
27. Wisz, M. S. et al. The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biol. Rev.* **88**, 15–30 (2013).
28. Mateo, R. G., Felicisimo, A. M., Pottier, J., Guisan, A. & Munoz, J. Do stacked species distribution models reflect altitudinal diversity patterns? *PLoS ONE* **7**, e32586 (2012).
29. Peterson, A. T., Navarro-Siguenza, A. G. & Benitez-Diaz, H. The need for continued scientific collecting; a geographic analysis of Mexican bird specimens. *Ibis* **140**, 288–294 (1998).
30. Hirzel, A. & Guisan, A. Which is the optimal sampling strategy for habitat suitability modelling. *Ecol. Modell.* **157**, 331–341 (2002).
31. Albert, C. H., Graham, C. H., Yoccoz, N. G., Zimmermann, N. E. & Thuiller, W. Applied sampling in ecology and evolution - integrating questions and designs. *Ecography* **33**, 1028–1037 (2010).
32. Leitao, P. J., Moreira, F. & Osborne, P. E. Effects of geographical data sampling bias on habitat models of species distributions: a case study with steppe birds in southern Portugal. *Int. J. Geogr. Inform. Sci.* **25**, 439–453 (2011).
33. Tessarolo, G., Rangel, T. F., Araujo, M. B. & Hortal, J. Uncertainty associated with survey design in species distribution models. *Divers. Distrib.* **20**, 1258–1269 (2014).
34. Vollering, J., Halvorsen, R., Auestad, I. & Rydgren, K. Bunching up the background betters bias in species distribution models. *Ecography* **42**, 1717–1727 (2019).
35. Tessarolo, G., Lobo, J. M., Rangel, T. F. & Hortal, J. High uncertainty in the effects of data characteristics on the performance of species distribution models. *Ecol. Indic.* **121**, 107147 (2021).
36. Graham, C. H. et al. The influence of spatial errors in species occurrence data used in distribution models. *J. Appl. Ecol.* **45**, 239–247 (2008).
37. Moudry, V. & Simova, P. Influence of positional accuracy, sample size and scale on modelling species distributions: a review. *Int. J. Geogr. Inform. Sci.* **26**, 2083–2095 (2012).
38. Hefley, T. J., Brost, B. M. & Hooten, M. B. Bias correction of bounded location errors in presence-only data. *Methods Ecol. Evol.* **8**, 1566–1573 (2017).
39. Guisan, A. & Rahbek, C. SESAM - a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *J. Biogeogr.* **38**, 1433–1444 (2011).
40. Jaeschke, A. et al. Biotic interactions in the face of climate change: a comparison of three modelling approaches. *PLoS ONE* **7**, e51472 (2012).
41. Dawson, M. N. et al. An horizon scan of biogeography. *Front. Biogeogr.* **5**, fb_18854 (2013).
42. Bruelheide, H. et al. sPlot - A new tool for global vegetation analyses. *J. Veg. Sci.* **30**, 161–186 (2019).
43. Sabatini, F. M. et al. sPlotOpen—An environmentally balanced, open-access, global dataset of vegetation plots. *Glob. Ecol. Biogeogr.* **30**, 1740–1764 (2021).
44. Zizka, A. et al. CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods Ecol. Evol.* **10**, 744–751 (2019).
45. Anderson, R. P. et al. Optimizing biodiversity informatics to improve information flow, data quality, and utility for science and society. *Front. Biogeogr.* **12**, e47839 (2020).
46. Grattarola, F., Bowler, D. & Keil, P. Integrating presence-only and presence-absence data to model changes in species geographic ranges: An example of yaguarundí in Latin America. Preprint available at EcoEvorxiv: https://doi.org/10.32942/osf.io/67c4u (2022).
47. Ficetola, G. F. et al. An evaluation of the robustness of global amphibian range maps. *J. Biogeogr.* **41**, 211–221 (2014).
48. Williams, K. J., Belbin, L., Austin, M. P., Stein, J. L. & Ferrier, S. Which environmental variables should I use in my biodiversity model? *Int. J. Geogr. Inform. Sci.* **26**, 2009–2047 (2012).
49. Lomolino, M.V. Conservation biogeography. In *Frontiers of biogeography: new directions in the geography of nature* (eds. Lomolino, M.V., Heaney, L.R) 293–296 (Sinauer Associates, Sunderland, MA, 2004).
50. Kuper, W., Sommer, J. H., Lovett, J. C. & Barthlott, W. Deficiency in African plant distribution data—missing pieces of the puzzle. *Botanical J. Linnean Soc.* **150**, 355–368 (2006).
51. Duputie, A., Zimmermann, N. E. & Chuine, I. Where are the wild things? Why we need better data on species distribution. *Glob. Ecol. Biogeogr.* **23**, 457–467 (2014).
52. Sousa-Baena, M. S., Garcia, L. C. & Peterson, A. T. Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Divers. Distrib.* **20**, 369–381 (2014).
53. Meyer, C., Kreft, H., Guralnick, R. & Jetz, W. Global priorities for an effective information basis of biodiversity distributions. *Nat. Commun.* **6**, 8221 (2015).
54. Wuest, R. O. et al. Macroecology in the age of Big Data - Where to go from here? *J. Biogeogr.* **47**, 1–12 (2020).
55. Dennis, R. L. H., Sparks, T. H. & Hardy, P. B. Bias in butterfly distribution maps: the effects of sampling effort. *J. Insect Conserv.* **3**, 33–42 (1999).
56. Hortal, J., Jimenez-Valverde, J., Gomez, J. F., Lobo, J. M. & Baselga, A. Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos* **117**, 847–858 (2018).
57. Kéry, M. Towards the modelling of true species distributions. *J. Biogeogr.* **38**, 617–618 (2011).
58. Gaiji, S. et al. Content assessment of the primary biodiversity data published through GBIF network: status, challenges and potentials. *Biodivers. Inform.* **8**, 94–172 (2013).
59. Hortal, J. et al. Seven shortfalls that beset large-scale knowledge of biodiversity. *Ann. Rev. Ecol. Evol. Syst.* **46**, 523–549 (2015).
60. Anderson, R.P. et al. Final report of the task group on GBIF data fitness for use in distribution modelling. Global Biodiversity Information Facility. 1-27(2016).

61. Girardello, M. et al. Gaps in butterfly inventory data: a global analysis. *Biol. Conserv.* **236**, 289–295 (2019).
62. Moudrý, V. & Devillers, R. Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data. *Ecol. Inform.* **56**, 101051 (2020).
63. Hughes, A. C. et al. Sampling biases shape our view of the natural world. *Ecography* **44**, 1259–1269 (2021).
64. Raja, N. B. et al. Colonial history and global economics distort our understanding of deep-time biodiversity. *Nat. Ecol. Evol.* **6**, 145–154 (2022).
65. Higgins, S. I. et al. A physiological analogy of the niche for projecting the potential distribution of plants. *J. Biogeogr.* **39**, 2132–2145 (2012).
66. Owens, H. L. et al. Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas. *Ecol. Modell.* **263**, 10–18 (2013).
67. Yates, K. L. et al. Outstanding challenges in the transferability of ecological models. *Trend. Ecol. Evol.* **33**, 790–802 (2018).
68. Qiao, H. et al. An evaluation of transferability of ecological niche models. *Ecography* **42**, 521–534 (2019).
69. Stohlgren, T. J., Jarnevich, C. S., Esaias, W. E. & Morisette, J. T. Bounding species distribution models. *Curr. Zool.* **57**, 642–647 (2011).
70. Mesgaran, M. B., Cousens, R. D. & Webber, B. L. Here be dragons: a tool for quantifying novelty due to covariate range and correlation change when projecting species distribution models. *Divers. Distrib.* **20**, 1147–1159 (2014).
71. Meyer, H. & Pebesma, E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods Ecol. Evol.* **12**, 1620–1633 (2021).
72. Shcheglovitova, M. & Anderson, R. P. Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sampl6e sizes. *Ecol. Modell.* **269**, 9–17 (2013).
73. De Oliveira, G., Rangel, T. F., Lima-Ribeiro, M. S., Terribile, L. C. & Diniz-Filho, J. A. F. Evaluating, partitioning, and mapping the spatial autocorrelation component in ecological niche modeling: a new approach based on environmentally equidistant records. *Ecography* **37**, 637–647 (2014).
74. Ploton, P. et al. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nat. Commun.* **11**, 4540 (2020).
75. Tessarolo, G., Ladle, R. J., Lobo, J. M., Rangel, T. F. & Hortal, J. Using maps of biogeographical ignorance to reveal the uncertainty in distributional data hidden in species distribution models. *Ecography* **44**, 1743–1755 (2021).
76. Thibaud, E., Petitpierre, B., Broennimann, O., Davison, A. C. & Guisan, A. Measuring the relative effect of factors affecting species distribution model predictions. *Methods Ecol. Evol.* **5**, 947–955 (2014).
77. Chevalier, M. et al. Low spatial autocorrelation in mountain biodiversity data and model residuals. *Ecosphere* **12**, e03403 (2021).
78. Meyer, H. & Pebesma, E. Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nat. Commun.* **13**, 2208 (2022).
79. Bruelheide, H. et al. Global trait-environment relationships of plant communities. *Nat. Ecol. Evol* **2**, 1906–1917 (2018).
80. Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B. & Schigel, D. Data integration enables global biodiversity synthesis. *Proc. Natl Acad. Sci.* **118**, e2018093118 (2021).
81. Maldonado, C. et al. Species diversity and distribution in the era of Big Data. *Glob. Ecol. Biogeogr.* **24**, 973–984 (2015).
82. Troudet, J. et al. Taxonomic bias in biodiversity data and societal preferences. *Sci. Rep.* **7**, 9132 (2017).
83. Nunez, M. A. & Amano, T. Monolingual searches can limit and bias results in global literature reviews. *Nat. Ecol. Evol.* **5**, 264 (2021a).
84. Nunez, M. A., Chiuffo, M. C., Pauchard, A. & Zenni, R. D. Making ecology really global. *Trend. Ecol. Evol.* **36**, 766–769 (2021b).
85. Adamo, M. et al. Plant scientists' research attention is skewed towards colourful, conspicuous and broadly distributed flowers. *Nat. Plants* **7**, 574–578 (2021).
86. Sanchez-Fernandez, D. et al. Don't forget subterranean ecosystems in climate change agendas. *Nat. Climate Change* **11**, 458–459 (2021).
87. Bini, L. M., Diniz-Filho, J. A. F., Rangel, T. F. L., Bastos, R. P. & Pinto, M. P. Challenging Wallacean and Linnean shortfalls: knowledge gradients and conservation planning in a biodiversity hotspot. *Divers. Distrib.* **12**, 475–482 (2006).
88. Oliver, R. Y., Meyer, C., Ranipeta, A., Winner, K. & Jetz, W. Global and national trends, gaps, and opportunities in documenting and monitoring species distributions. *PLoS Biol.* **19**, e3001336 (2021).
89. Sastre, P. & Lobo, J. M. Taxonomist survey biases and the unveiling of biodiversity patterns. *Biol. Conserv.* **142**, 462–467 (2009).
90. Boakes, E. H. et al. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biol.* **8**, e1000385 (2010).
91. Yang, W., Ma, K. & Kreft, H. Geographical sampling bias in a large distributional database and its effects on species richness-environment models. *J. Biogeogr.* **40**, 1415–1426 (2013).
92. Kadmon, R., Farber, O. & Danin, A. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecol. Appl.* **14**, 401–413 (2004).
93. Oliveira, U. et al. The strong influence of collection bias on biodiversity knowledge shortfalls of Brazilian terrestrial biodiversity. *Divers. Distrib.* **22**, 1232–1244 (2016).
94. Geldmann, J. et al. What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. *Divers. Distrib.* **22**, 1139–1149 (2016).
95. Ronquillo, C. et al. Assessing spatial and temporal biases and gaps in the publicly available distributional information of iberian mosses. *Biodivers. Data J.* **8**, e53474 (2020).
96. Petersen, T. K., Speed, J. D. M., Grotan, V. & Austrheim, G. Species data for understanding biodiversity dynamics: The what, where and when of species occurrence data collection. *Ecol. Solut. Evid.* **2**, e12048 (2021).
97. Pärtel, M., Sabatini, F. M., Morueta-Holme, N., Kreft, H. & Dengler, J. Macroecology of vegetation - Lessons learnt from the Virtual Special Issue. *J. Veg. Sci.* **33**, e13121 (2022).
98. Rodrigues, A. S. L. et al. A global assessment of amphibian taxonomic effort and expertise. *Bioscience* **60**, 798–806 (2010).
99. Meyer, C., Weigelt, P. & Kreft, H. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecol. Lett.* **19**, 992–1006 (2016).
100. Costa, G. C., Nogueira, C., Machado, R. B. & Colli, G. R. Sampling bias and the use of ecological niche modeling in conservation planning: a field evaluation in a biodiversity hotspot. *Biodivers. Conserv.* **19**, 883–899 (2010).
101. Rocchini, D. et al. Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. *Progr. Phys. Geogr.* **35**, 211–226 (2011).
102. Phillips, S. J. et al. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecol. Appl.* **19**, 181–197 (2009).
103. Beck, J., Boller, M., Erhardt, A. & Schwanghart, W. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecol. Inform.* **19**, 10–15 (2014).
104. Barbosa, A. M., Pautasso, M. & Figueiredo, D. Species-people correlations and the need to account for survey effort in biodiversity analyses. *Divers. Distrib.* **19**, 1188–1197 (2013).
105. Chevalier, M., Broennimann, O., Cornuault, J. & Guisan, A. Data integration methods to account for spatial niche truncation effects in regional projections of species distribution. *Ecol. Appl.* **31**, e02427 (2021).
106. Acevedo, P., Jimenez-Valverde, A., Lobo, J. M. & Real, R. Delimiting the geographical background in species distribution modelling. *J. Biogeogr.* **39**, 1383–1390 (2012).
107. Jimenez-Valverde, A., Acevedo, P., Barbosa, A. M., Lobo, J. M. & Real, R. Discrimination capacity in species distribution models depends on the representativeness of the environmental domain. *Glob. Ecol. Biogeogr* **22**, 508–516 (2013).
108. Sillero, N. & Barbosa, A. M. Common mistakes in ecological niche models. *Int. J. Geogr. Inform. Sci.* **35**, 213–226 (2021).
109. Sobral-Souza, T. et al. Knowledge gaps hamper understanding the relationship between fragmentation and biodiversity loss: the case of Atlantic Forest fruit-feeding butterflies. *PeerJ* **9**, e11673 (2021).
110. McCune, J. L., Rosner-Katz, H., Bennett, J. R., Schuster, R. & Kharouba, H. M. Do traits of plant species predict the efficacy of species distribution models for finding new occurrences? *Ecol. Evol.* **10**, 5001–5014 (2020).
111. Guo, C. et al. Uncertainty in ensemble modelling of large-scale species distribution: effects from species characteristics and model techniques. *Ecol. Modell.* **306**, 67–75 (2015).
112. Jimenez-Valverde, A., Lobo, J. M. & Hortal, J. Not as good as they seem: the importance of concepts in species distribution modelling. *Divers. Distrib.* **14**, 885–890 (2008).
113. Jeliazkov, A. et al. Sampling and modelling rare species: conceptual guidelines for the neglected majority. *Glob. Change Biol.* **28**, 3754–3777 (2022).
114. Anderson, R. P. & Raza, A. The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus Nephelomys) in Venezuela. *J. Biogeogr.* **37**, 1378–1393 (2010).
115. Hattab, T. et al. A unified framework to model the potential and realized distributions of invasive species within the invaded range. *Divers. Distrib.* **23**, 806–819 (2017).
116. Lembrechts, J. J., Lenoir, J., Scheffers, B. & De Frenne, P. Designing countrywide and regional microclimate networks. *Glob. Ecol. Biogeogr.* **30**, 1168–1174 (2021).
117. Fourcade, Y., Engler, J. O., Rodder, D. & Secondi, J. 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 (2014).
118. Nunez-Penichet, C. et al. Selection of sampling sites for biodiversity inventory: Effects of environmental and geographical considerations. *Methods Ecol. Evol* **13**, 1595–1607 (2022).
119. Whittaker, R. H. A criticism of the plant association and climatic climax concepts. *Northwest Sci.* **26**, 17–31 (1951).

120. Austin, M. P., Cunningham, R. B. & Fleming, P. M. New approaches to direct gradient analysis using environmental scalars and statistical curve-fitting procedures. *Vegetatio* **55**, 11–27 (1984).
121. Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B. & Anderson, R. P. spthin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography* **38**, 541–545 (2015).
122. Fourcade, Y. Fine-tuning niche models matters in invasion ecology.A lesson from the land planarian *Obama nungara.Ecol. Modell.* **457**, 109686 (2021).
123. Varela, S., Anderson, R. P., Garcia-Valdes, R. & Fernandez-Gonzalez, F. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography* **37**, 1084–1091 (2014).
124. Anderson, R. P. & Gonzalez Jr, I. Species-specific tuning increases robustness to sampling bias in models of species distributions: An implementation with Maxent. *Ecol. Modell.* **222**, 2796–2811 (2011).
125. Gabor, L., Moudry, V., Bartak, V. & Lecours, V. How do species and data characteristics affect species distribution models and when to use environmental filtering? *Int. J. Geogr. Inform. Sci.* **34**, 1567–1584 (2020).
126. Dormann, C. F. et al. Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecol. Monogr.* **88**, 485–504 (2018).
127. Hao, T., Elith, J., Guillera-Arroita, G. & Lahoz-Monfort, J. J. A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. *Divers. Distrib.* **25**, 839–852 (2019).
128. Hao, T., Elith, J., Lahoz-Monfort, J. J. & Guillera-Arroita, G. Testing whether ensemble modelling is advantageous for maximising predictive performance of species distribution models. *Ecography* **43**, 549–558 (2020).
129. Amano, T., Lamming, J. D. L. & Sutherland, W.-J. Spatial gaps in global biodiversity information and the role of citizen science. *Bioscience* **66**, 393–400 (2016).
130. Wolf, S. et al. Citizen science plant observations encode global trait patterns. *Nat. Ecol. Evol.* **6**, 1850–1859 (2022).
131. Theobald, E. J. et al. Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biol. Conserv.* **181**, 236–244 (2015).
132. Lobo, J. M. et al. KnowBR: An application to map the geographical variation of survey effort and identify well-surveyed areas from biodiversity databases. *Ecol. Indic.* **91**, 241–248 (2018).
133. Mokany, K., Harwood, T. D., Overton, J. M., Barker, G. M. & Ferrier, S. Combining alpha- and beta-diversity models to fill gaps in our knowledge of biodiversity. *Ecol. Lett.* **14**, 1043–1051 (2011).
134. Chevalier, M., Zarzo-Arias, A., Guélat, J., Mateo, R. G. & Guisan, A. Accounting for niche truncation to improve spatial and temporal predictions of species distributions. *Front. Ecol. Evol.* **10**, 944116 (2022).
135. Boggs, S. W. An atlas of ignorance: A needed stimulus to honest thinking and hard work. *Proc. Am. Philos. Soc.* **93**, 253–258 (1949).
136. Ladle, R. J. & Hortal, J. Mapping species distributions: living with uncertainty. *Front. Biogeogr.* **5**, 8–9 (2013).
137. Konig, C. et al. Biodiversity data integration-the significance of data resolution and domain. *PLoS Biol.* **17**, e3000183 (2019).
138. Ellis-Soto, D., Merow, C., Amatulli, G., Parra, J. L. & Jetz, W. Continental-scale 1 km hummingbird diversity derived from fusing point records with lateral and elevational expert information. *Ecography* **44**, 640–652 (2021).
139. Palmer, M. W. How should one count species? *Nat. Areas J.* **15**, 124–135 (1995).
140. Cazzolla Gatti, R. et al. The number of tree species on Earth. *Proc. Natl Acad. Sci.* **119**, e2115329119 (2022).
141. Soberón, J. M., Llorente, J. B. & Oñate, L. The use of specimen-label databases for conservation purposes: an example using Mexican Papilionid and Pierid butterflies. *Biodivers. Conserv.* **9**, 1441–1466 (2000).
142. Bystriakova, N., Peregrym, M., Erkens, R. H. J., Bezsmertna, O. & Schneider, H. Sampling bias in geographic and environmental space and its effect on the predictive power of species distribution models. *Syst. Biodivers.* **10**, 305–315 (2012).
143. Zurell, D. et al. The virtual ecologist approach: simulating data and observers. *Oikos* **119**, 622–635 (2010).
144. Miller, J. A. Virtual species distribution models: Using simulated data to evaluate aspects of model performance. *Progr. Phys. Geogr.* **38**, 117–128 (2014).
145. Moudry, V. Modelling species distributions with simulated virtual species. *J. Biogeogr.* **42**, 1365–1366 (2015).
146. Fernandes, R. F., Scherrer, D. & Guisan, A. How much should one sample to accurately predict the distribution of species assemblages? A virtual community approach. *Ecol. Inform.* **48**, 125–134 (2018).
147. Meynard, C. N., Leroy, B. & Kaplan, D. M. Testing methods in species distribution modelling using virtual species: what have we learnt and what are we missing? *Ecography* **42**, 2021–2036 (2019).
148. Dullinger, S. et al. Extinction debt of high-mountain plants under twenty-first-century climate change. *Nat. Climate Change* **2**, 619–622 (2012).
149. Pulliam, H. R. Sources, sinks, and population regulation. *Am. Nat.* **132**, 652–661 (1988).
150. Zurell, D. et al. Benchmarking novel approaches for modelling species range dynamics. *Glob. Change Biol.* **22**, 2651–2664 (2016).
151. Brun, P. et al. Model complexity affects species distribution projections under climate change. *J. Biogeogr.* **47**, 130–142 (2020).
152. Schweiger, A. H., Irl, S. D. H., Steinbauer, M. J., Dengler, J. & Beierkuhnlein, C. Optimizing sampling approaches along ecological gradients. *Methods Ecol. Evol.* **7**, 463–471 (2016).
153. Meynard, C. N. & Kaplan, D. M. Using virtual species to study species distributions and model performance. *J. Biogeogr.* **40**, 1–8 (2013).
154. Hirzel, A. H., Helfer, V. & Metral, F. Assessing habitat-suitability models with a virtual species. *Ecol. Modell.* **145**, 111–121 (2011).
155. Leroy, B., Meynard, C. N., Bellard, C. & Courchamp, F. virtualspecies, an R package to generate virtual species distributions. *Ecography* **39**, 599–607 (2016).
156. Qiao, H. et al. NicheA: creating virtual species and ecological niches in multivariate environmental scenarios. *Ecography* **39**, 805–813 (2016).
157. De Paor, D. G. & Whitmeyer, S. J. Geological and geophysical modeling on virtual globes using KML, COLLADA, and Javascript. *Comput. Geosci.* **37**, 100–110 (2011).
158. Wallace, A. R. On the law which has regulated the introduction of new species. *Ann. Mag. Nat. Hist.* **16**, 184–196 (1855).
159. Pillar, V. D., Sabatini, F. M., Jandt, U., Camiz, S. & Bruelheide, H. Revealing the functional traits linked to hidden environmental factors in community assembly. *J. Veg. Sci.* **32**, e12976 (2011).
160. Cazzolla Gatti, R. A century of biodiversity: Some open questions and some answers. *Biodiversity* **18**, 175–185 (2017).
161. Qiao, H., Peterson, A. T., Ji, L. & Hu, J. Using data from related species to overcome spatial sampling bias and associated limitations in ecological niche modelling. *Methods Ecol. Evol.* **8**, 1804–1812 (2017).
162. Winsberg, E. Sanctioning models: The epistemology of simulation. *Sci. Context* **12**, 275–292 (1999).
163. Winsberg, E. Simulated experiments: Methodology for a virtual world. *Philos. Sci.* **70**, 105–125 (2003).
164. Peck, S. L. Simulation as experiment: a philosophical reassessment for biological modeling. *Trend Ecol. Evol.* **19**, 530–534 (2004).
165. Rangel, T. F. L., Diniz-Filho, J. A. F. & Colwell, R. K. Species richness and evolutionary niche dynamics: a spatial pattern-oriented simulation experiment. *Am. Nat.* **170**, 602–616 (2007).
166. Nakazawa, Y. Niche breadth, environmental landscape, and physical barriers: their importance as determinants of species distributions. *Biol. J. Linnean Soc.* **108**, 241–250 (2013).
167. Nakazawa, Y. & Peterson, A. T. Effects of climate history and environmental grain on species' distributions in Africa and South America. *Biotropica* **47**, 292–299 (2015).
168. Darroch, S. A. & Saupe, E. E. Reconstructing geographic range-size dynamics from fossil data. *Paleobiology* **44**, 25–39 (2018).
169. Kadmon, R., Farber, O. & Danin, A. A systematic analysis of factors affecting the performance of climatic envelope models. *Ecol. Appl.* **13**, 853–867 (2003).
170. Foody, G. M. GIS: stressing the geographical. *Progr. Phys. Geogr.* **28**, 152–158 (2004).
171. Osborne, P. E., Foody, G. M. & Suarez-Seoane, S. Non-stationarity and local approaches to modelling the distributions of wildlife. *Divers. Distrib.* **13**, 313–323 (2007).
172. Foody, G. M. GIS: biodiversity applications. *Progr. Phys. Geogr.* **32**, 223–235 (2008).
173. Hortal, J., Lobo, J. M. & Jimenez-Valverde, A. Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife (Canary Islands). *Conserv. Biol.* **21**, 853–863 (2007).
174. Ficetola, G. F. et al. Sampling bias inverts ecogeographical relationships in island reptiles. *Glob. Ecol. Biogeogr.* **23**, 1303–1313 (2014).
175. Rocchini, D. et al. Anticipating species distributions: handling sampling effort bias under a Bayesian framework. *Sci. Total Environ.* **584–585**, 282–290 (2017).
176. Miller, D. L., Burt, M. L., Rexstad, E. A. & Thomas, L. Spatial models for distance sampling data: recent developments and future directions. *Methods Ecolo. Evol.* **4**, 1001–1010 (2013).
177. Barry, S. C. & Welsh, A. H. Generalized additive modelling and zero inflated count data. *Ecol. Modell.* **157**, 179–188 (2002).
178. Schmera, D. & Eros, T. The role of sampling effort, taxonomical resolution and abundance weight in multivariate comparison of stream dwelling caddisfly assemblages collected from riffle and pool habitats. *Ecol. Indic.* **11**, 230–239 (2011).
179. Lark, R. M. Spatially nested sampling schemes for spatial variance components: scope for their optimization. *Comput. Geosci.* **37**, 1633–1641 (2011).
180. Lark, R. M. Exploring scale-dependent correlation of soil properties by nested sampling. *Eur. J. Soil Sci.* **56**, 307–317 (2005).

181. Rousset, F. & Ferdy, J.-B. Testing environmental and genetic effects in the presence of spatial autocorrelation. *Ecography* **37**, 781–790 (2014).
182. Rousset, F., Ferdy, J.-B. and Courtiol, A. spaMM: Mixed-Effect Models, with or without Spatial Random Effects. R package version 3.11.14. (2021). https://CRAN.R-project.org/package=spaMM.
183. Lozier, J. D., Aniello, P. & Hickerson, M. J. Predicting the distribution of Sasquatch in western North America: anything goes with ecological niche modelling. *J. Biogeogr.* **36**, 1623–1627 (2009).
184. Tessarolo, G., Ladle, R., Rangel, T. & Hortal, J. Temporal degradation of data limits biodiversity research. *Ecol. Evol.* **7**, 6863–6870 (2017).
185. Foody, G. M. Impacts of imperfect reference data on the apparent accuracy of species presence-absence models and their predictions. *Global Ecol. Biogeogr.* **20**, 498–508 (2011).
186. Beale, C. M. & Lennon, J. J. Incorporating uncertainty in predictive species distribution modelling. *Philos. Transac. R. Soc. B* **367**, 247–258 (2012).
187. Sanchez-Fernandez, D., Lobo, J. M., Abellan, P., Ribera, I. & Millan, A. Bias in freshwater biodiversity sampling: the case of Iberian water beetles. *Divers. Distrib.* **14**, 754–762 (2008).
188. Gomez-Rodriguez, C., Bustamante, J., Díaz-Paniagua, C. & Guisan, A. Integrating detection probabilities in species distribution models of amphibians breeding in Mediterranean temporary ponds. *Divers. Distrib.* **18**, 260–272 (2012).
189. Ficetola, G. F., Bonardi, A., Sindaco, R. & Padoa-Schioppa, E. Estimating patterns of reptile biodiversity in remote regions. *J. Biogeogr.* **40**, 1202–1211 (2013).
190. Barbosa, A. M., Fontaneto, D., Marini, L. & Pautasso, M. Is the human population a large-scale indicator of the species richness of ground beetles? *Animal Conserv.* **13**, 432–441 (2010).
191. Fontaneto, D., Barbosa, A. M., Segers, H. & Pautasso, M. The 'rotiferologist' effect and other global correlates of species richness in monogonont rotifers. *Ecography* **35**, 174–182 (2012).
192. Real, R., Barbosa, A. M. & Bull, J. W. Species Distributions, quantum theory, and the enhancement of biodiversity measures. *Syst. Biol.* **66**, 453–462 (2017).
193. Wasof, S. et al. Disjunct populations of European vascular plant species keep the same climatic niches. *Glob. Ecol. Biogeogr.* **24**, 1401–1412 (2015).
194. McCarthy, M. A. & Masters, P. Profiting from prior information in Bayesian analyses of ecological data. *J. Appl. Ecol.* **42**, 1012–1019 (2005).
195. Elith, J., Burgman, M. A. & Regan, H. M. Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecol. Modell.* **157**, 313–329 (2002).
196. Radosavljevic, A. & Anderson, R. P. Making better Maxent models of species distributions: complexity, overfitting and evaluation. *J. Biogeogr.* **41**, 629–643 (2014).
197. Rosner-Katz, H., McCune, J. L. & Bennett, J. R. Using stacked SDMs with accuracy and rarity weighting to optimize surveys for rare plant species. *Biodivers. Conserv.* **29**, 3209–3225 (2020).
198. Zurell, D., Elith, J. & Schroder, D. Predicting to new environments: tools for visualizing model behaviour and impacts on mapped distributions. *Divers. Distrib.* **18**, 628–634 (2012).
199. Brooks, T. M. et al. Coverage provided by the global protected-area system: is it enough? *BioScience* **54**, 1081–1091 (2004).
200. Dorling, D. Area Cartograms: Their Use and Creation. Concepts and Techniques in Modern Geography (CATMOG) 59 (Univ. of East Anglia, Norwich, U.K.) (1996).
201. Gastner, M. T. & Newman, M. E. J. Diffusion-based method for producing density-equalizing maps. *Proc. Natl Acad. Sci.* **101**, 7499–7504 (2004).

ACKNOWLEDGEMENTS

We are grateful to the handling Editor and to three anonymous reviewers for precious suggestions on a previous draft of the paper. Arianna Ferrara partially edited Fig. 1 of this manuscript.

This study has received funding from the project SHOWCASE (SHOWCASing synergies between agriculture, biodiversity and ecosystems services to help farmers capitalising on native biodiversity) within the European Union's Horizon 2020 Researcher and Innovation Programme under grant agreement No. 862480. D.R. was partially funded by a research project implemented under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4 - Call for tender No. 3138 of 16 December 2021, rectified by Decree n.3175 of 18 December 2021 of Italian Ministry of University and Research funded by the European Union - NextGenerationEU. Project code CN_00000033, Concession Decree No. 1034 of 17 June 2022 adopted by the Italian Ministry of University and Research, CUP J33C22001190001, Project title "National Biodiversity Future Center - NBFC". D.R. was also patially funded by the Horizon Europe projects Earthbridge and B^3. F.M.S. gratefully acknowledges financial support by the Rita-Levi Montalcini (2019) programme, funded by the Italian Ministry of University. E.T. is supported by the Estonian Research Council (grant code MOBJD1030).

AUTHOR CONTRIBUTIONS

D.R., E.T., E.M., M.M. and P.Z. analyzed the presented data. All authors worked on the conceptualization and the development of the manuscript during the writing phase.

COMPETING INTERESTS

J.H. is Editor-in-Chief of npj Biodiversity. All other auuthors declare having no competing interests as defined by Nature Portfolio, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44185-023-00014-6.

Correspondence and requests for materials should be addressed to Duccio Rocchini.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.