Causal inference and large-scale expert validation shed light on the drivers of SDM accuracy and variance

Robin J. Boyd[1], Martin Harvey[1], David B. Roy[1], Tony Barber[10], Karen A. Haysom[7], Craig R. Macadam[2,5], Roger K.A. Morris[2,3], Carolyn Palmer[9], Stephen Palmer[9], Chris D. Preston[6], Pam Taylor[8], Rob Ward[7], Stuart G. Ball[4], Oliver L. Pescott[1]

[1]UK Centre for Ecology and Hydrology, Benson Ln, Crowmarsh Gifford, Oxfordshire, UK.

[2]Scientific Associate, The Natural History Museum, Cromwell Rd, South Kensington, London SW7 5BD

[3]Hoverfly Recording Scheme, 241 Commonside East, Mitcham, Surrey

[4]Hoverfly Recording Scheme, 255 Eastfield Road, Peterborough PE1 4BH

[5]Riverfly Recording Schemes, c/o Buglife Scotland, Unit 4, Beta Centre, Stirling University Innovation Park, Stirling FK9 4HN.

[6]19 Green's Road, Cambridge CB4 3EF

[7]Amphibian and Reptile Conservation, 744 Christchurch Road, Boscombe, Bournemouth, Dorset, BH7 6BZ

[8]British Dragonfly Society, (registered address) Ashcroft, Brington Road, Old Weston, Huntingdon, PE28 5LP

[9]Gelechiid Recording Scheme, 137 Lightfoot Lane, Fulwood, Preston, Lancs PR4 0AH, England

[10]British Myriapod and Isopod Group.

Running title: Causes of interspecies variation in SDM performance

Abstract

1. The literature is awash with studies purporting to show how various species and data characteristics affect the performances of Species Distribution Models (SDMs). Many of these studies follow a similar template: they fit SDMs for several species, or the same species using different datasets; assess the accuracy of those SDMs using skill statistics; and then identify correlates thereof. Interpreting the findings of these studies is challenging because skill statistics can reflect species and data characteristics rather than model accuracy, and correlates of model performance are not necessarily causes.
2. Here, we took a different approach to identifying causes of variation in SDM performance. We fitted models for 535 species across 5 invertebrate groups and 1 plant group in the United Kingdom (UK), using a fairly typical SDM workflow. We measured two components of SDM performance: the variance among replicate fits and accuracy. Rather than using skill statistics, accuracy was assessed by taxon experts. We constructed Directed Acyclic Graphs (DAGs)

depicting plausible effects of explanatory variables (e.g. species' prevalence, sample size) on SDM performance, then quantified those effects using multilevel piecewise path models.
3. We found that the degree to which the available data covered species' environmental niches was the only explanatory variable to affect SDM accuracy. We suggest that previously reported associations between sample size and SDM accuracy reflect improved coverage of species environmental niches at higher sample sizes; that is to say, niche completeness confounds the effect of sample size on SDM accuracy. We also report that the completeness of species' environmental niches, sample size, species' prevalence and the degree to which the available data cover species' geographic ranges affect SDM variance.
4. Our results demonstrate the challenges associated with the high-throughput approach to modelling species' distributions. There is no guarantee that accurate and precise SDMs can be constructed for large numbers of species unless their ranges and niches have been sampled comprehensively. Decisions about whether modelling is worthwhile should not be based on simple criteria like sample size.

Keywords: Directed Acyclic Graph; Structural Equation Modelling; causal inference; expert elicitation; species distribution modelling

Introduction

Species Distribution Models (SDMs, also known as habitat suitability models) estimate species' environmental preferences. Put very simply, they do so by comparing the environment at locations where a species was observed with the environment at locations where it was not. Once this comparison has been made, the SDM can be used to predict habitat suitability at any geographic location and point in time for which the relevant environmental data are available. This feature of SDMs makes them extremely useful for such applications as predicting the spread of invasive species (Montalva et al., 2017) and disease vectors (Simons et al., 2019), predicting future shifts in species' distributions in response to climate change (Stewart et al., 2022), and spatial conservation planning (El-Gabbas et al., 2020). If SDMs are to be applied in such settings, however, it is important that they perform well in terms of predicting habitat suitability.

The performance of a SDM may be decomposed broadly into its accuracy and precision (Bazzichetto et al., 2022). Accuracy is a measure of how close the model's predictions are to the "truth" on average. The most commonly used measure of a SDM's accuracy is its discrimination ability, that is, its ability to predict higher habitat suitability at locations where the species was observed than locations where it was not (Jiménez-valverde et al., 2013). Precision, on the other hand, is a measure of the variability among predictions from replicate model fits, which might include variability among SDM algorithms where multi-model ensembles are constructed (Watling et al., 2015). Models with high accuracy and precision will consistently make predictions that are close to the truth; clearly, it is desirable to know the situations in which this can be expected.

The literature is awash with studies purporting to show how various methodological decisions, data characteristics and species traits affect SDM performance. Methodological decisions include the choice of SDM algorithm or ensemble of algorithms (Fukuda & De Baets, 2016; Hao et al., 2020), environmental covariates (Arenas-Castro et al., 2022; Bucklin et al., 2015; De Marco & Nóbrega, 2018), and strategies to mitigate undesirable properties of the occurrence data (Barbet-Massin et al., 2012; Beck et al., 2014; Chapman et al., 2019; Dudík et al., 2005; Fourcade et al., 2014; Phillips et al., 2009). Data characteristics include the extent of spatial clustering and geographic bias (Bazzichetto et al., 2022; Beck et al., 2014; Steen et al., 2020), the expertise of data collectors (Steen et al., 2019), the ratio of presences to absences (Fukuda & De Baets, 2016), coverage of species'

geographic ranges (Konowalik & Nosol, 2021), and sample size (Feeley & Silman, 2011; Hernandez et al., 2006; Stockwell & Peterson, 2002; Wisz et al., 2008). And finally, species traits include range size relative to the study extent (Santika, 2011) and niche breadth (Hernandez et al., 2006; Tessarolo et al., 2021), amongst others. Most of the studies listed above follow a similar template: they fit SDMs for several species, or for the same species using different methodologies and datasets, then assess the accuracy of those models.

Assessing the accuracy of a SDM generally involves comparing its predictions to data. This data might be the same data that was used for model fitting, data withheld when fitting the model, or completely independent data (e.g. from a separate survey). Alternatively, in simulation studies, where virtual species are used, SDM predictions can be compared to those species' true distributions directly. Regardless, predictive accuracy is typically evaluated using skill statistics, such as the Area Under the receiver operator Curve (AUC), the True Skill Statistic (TSS) and Cohen's Kappa (Allouche et al., 2006; Leroy et al., 2018).

Although widely-used, skill statistics have been criticised on several grounds. A major limitation is that they depend on the focal species' prevalence, that is, its range size relative to the study extent (Jiménez-valverde et al., 2013; Lobo et al., 2008). This is problematic even for simulation studies, where the species' true distribution is known. Another problem is that, where sample prevalence—i.e. the ratio of presences to absences in the evaluation data—does not equal the species' true prevalence, skill statistics may not reflect a model's discrimination ability (Leroy et al., 2018). This limitation is particularly problematic given the widespread use of pseudo-absences in place of true absences, and the fact that true prevalence is seldom known. A related issue is that skill statistics give equal weight to presences and (pseudo) absences by virtue of their mathematical formulations, despite the fact that pseudo-absences are not observed. Notwithstanding these limitations and others (e.g. Lobo et al., 2010), AUC, kappa and TSS remain the most common measures of SDM accuracy.

Whilst most studies evaluate SDM accuracy using skill statistics, an alternative is to solicit expert opinion. For example, Smart et al. (2019) sought expert opinion on the realism of species response curves estimated by small-scale niche models for vascular plants and bryophytes in the United Kingdom (UK). Similarly, Beck et al. (2014) sought expert opinion on the spatial predictions produced by various SDMs for a European butterfly. These latter authors found that model accuracy increased when the occurrence data were thinned to reduce spatial clustering. However, this finding was evident only to the expert: it was not reflected by an increase in AUC. This clearly demonstrates that expert validation can, at the very least, provide a different perspective to skill statistics on what determines SDM accuracy.

Whether using expert opinion or skill statistics, appropriately quantifying SDM performance is only the first step towards understanding its determinants. The researcher must then quantify the relationships between the performance measures and predictors thereof. This is often achieved using some form of regression analysis—e.g. multiple regression, partial regression, ANOVA or t-tests (Barbet-Massin et al., 2012; De Marco & Nóbrega, 2018; Feeley & Silman, 2011; Steen et al., 2019; Tessarolo et al., 2021; Watling et al., 2015; Wisz et al., 2008)—or even simpler measures of correlation (Hernandez et al., 2006).

Whilst clearly useful, regression does not necessarily tell the full story when it comes to ascertaining the effects of independent variables on a response variable. It is well known that regression coefficients vary as independent variables are added to and removed from the model (Angrist & Pischke, 2009). Indeed, using regression for causal inference requires assumptions about all

confounders having been measured and included in the model (Gelman & Hill, 2006; McElreath, 2020). Furthermore, as it is typically used—i.e. with one response variable—regression cannot deal with indirect effects, which occur where one variable mediates the effect of a second variable on the response (Baron & Kenny, 1986).

In other disciplines, and to a lesser extent in ecology (but see Grace, 2006), the limitations of regression mentioned above have been long recognised and overcome using graph theory and causal analysis. Directed Acyclic Graphs (DAGs; Greenland et al., 1999; Pearl et al., 2016) are constructed to codify researchers' theories about how the explanatory variables affect the response variable(s). DAGS might reveal confounders that must be included in a regression analysis in order to produce unbiased coefficients. They might also reveal mediation pathways, or multiple response variables; in this case, path analysis, or more complex structural equation models, can be used to estimate the effects of interest (Grace, 2006).

Here, we used graph theory, causal analysis and expert validation to understand the drivers of SDM performance. We fitted SDMs for 1216 species of insect and bryophyte in the United Kingdom (UK), using a fairly typical presence/pseudo-absence modelling workflow. We evaluated the performances of a subset (535; 44%) of these models, both in terms of variance among replicate model fits, and accuracy as assessed by taxon experts. We used DAGs to conceive plausible models describing the effects of explanatory variables on SDM performance. We then used multilevel path analysis to quantify those effects, given our SDM workflow.

Methodology

Species occurrence data

We fitted SDMs using presence-only species occurrence records. The data were supplied by national recording schemes in the UK, who collate records made by volunteer expert naturalists for their taxon group of interest. For most taxa, we used the same data as Outhwaite et al. (2019), but applied additional filters. We only used gridded records collected at 1 km^2 or finer between 2000 and 2015 to match the SDM covariate data (supplementary material one), and removed records that were duplicated in terms of grid cell and species (standard practice for species distribution modelling).

Species Distribution Models

In this section, we briefly outline the SDM workflow (Fig. 1), but refer the reader to the ODMAP (Overview, Data, Model, Assessment and Prediction; Zurell et al., 2020) document in supplementary material one for full details. We used three SDM algorithms to estimate species' habitat suitability: Maxent, regularized logistic GLMs and random forests. We used the species occurrence data outlined above, and pseudo-absences generated according to the "non-overlapping target group" approach (Cerasoli et al., 2017; Phillips et al., 2009), as response variables. Twenty-five topographic, land cover and climate variables were used as covariates. We split the data randomly into five equally-sized subsets, then fitted each algorithm five times, leaving out one subset each time. Hence, we fitted 15 models for each species, which enabled us to assess the variability among replicate fits. The models were fitted at a spatial resolution of 1 km^2 on the British Ordnance Survey grid (EPSG:27700). Ensemble predictions were generated for each species by taking a weighted average (based on AUC) of the fifteen replicate model fits (Boyd et al., 2022).

We use the R (R Core Team, 2019) package soaR (https://github.com/robboyd/soaR) to fit, average and evaluate the models. soaR wraps around the packages glmnet, randomForest (Breiman et al., 2018) and dismo (Hijmans et al., 2017).

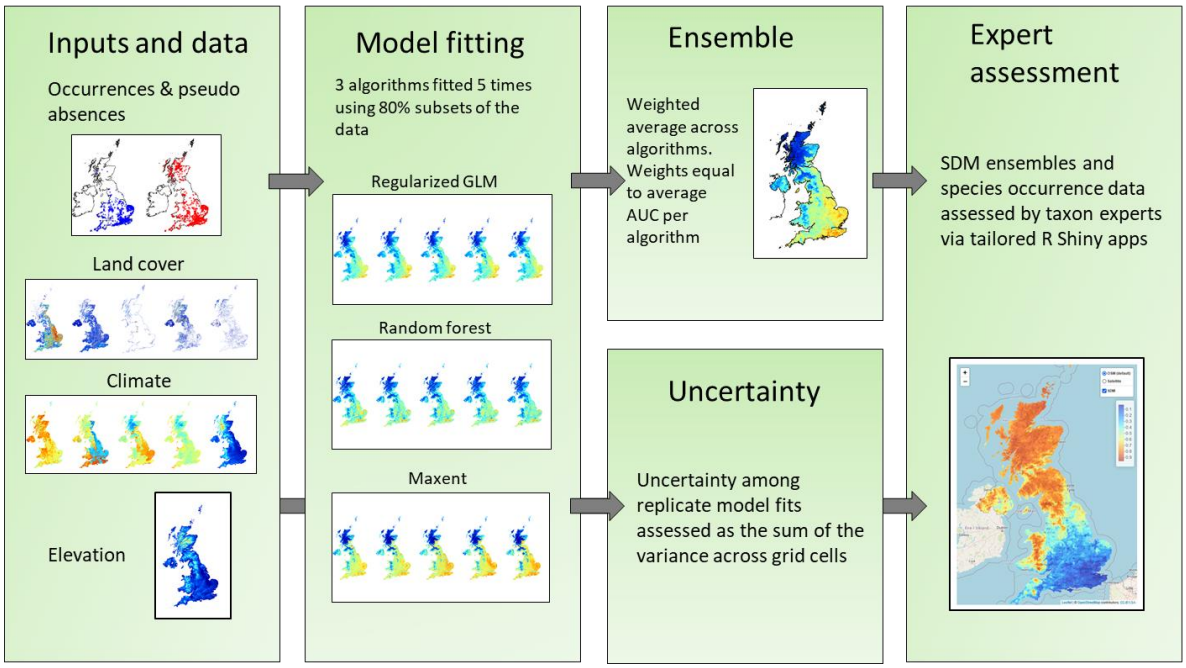

Figure 1. Species distribution modelling and assessment workflow. See the supplementary ODMAP document for full details (SM2).

Expert assessments of SDMs and data

Taxon experts (Table 1) assessed the available records and ensemble SDM predictions, in geographic space, for all species in their group of interest (or a random subset of 100 species in the case of the more speciose bryophytes; Table 1). Amongst other questions, they were asked 1) whether the available records for each species cover its environmental niche; 2) whether the available records for each species cover its geographic range; and 3) whether the map of predicted habitat suitability for each species (i.e. the ensemble SDM) reflects its true environmental niche in geographic space. The experts provided their answers to these questions on Likert scales ranging from 1 (excellent coverage/excellent habitat suitability predictions) to 5 (extremely poor coverage/extremely poor habitat suitability predictions).

Each expert was provided with a tailored R Shiny app, which included the predicted maps of habitat suitability, a map of the records used to fit the SDMs, maps of the environmental layers used to fit the models, and various questions including those listed above. Example code, containing all of these questions, can be found in (Pescott, 2022).

Table 1. A taxonomic breakdown of the number of species modelled, the number of models assessed, the assessors initials (see author list) and their affiliations.

Taxonomic group	**Number of species modelled**	**Number of species assessed**	**Expert initials**	**Recording scheme**
Mosses, liverworts and hornworts (Bryophyta, Marchantiophyta, and Anthocerotophyta)	782	100	CDP	British Bryological Society (https://www.britishbryologicalsociety.org.uk/)
Centipedes (Chilopoda)	29	29	TB	British Myriapod and Isopod Group, Centipede

				Recording Scheme (https://www.bmig.org.uk/)
Dragonflies (Odonta)	46	46	PT	British Dragonfly Society Recording Scheme (https://british-dragonflies.org.uk/)
Hoverflies (Syrphidae)	226	226	RM	Dipterists Forum, Hoverfly Recording Scheme (http://hoverfly.uk/hrs/)
Mayflies (Ephemeroptera)	38	38	CM	Riverfly Recording Schemes: Ephemeroptera (http://www.ephemeroptera.org.uk/)
Soldierflies and allies (Lower Brachycera)	95	95	MH	Soldierflies and Allies Recording Scheme (http://soldierflies.brc.ac.uk/)

Conceptual models describing SDM performance

Measures of SDM performance

We considered two distinct aspects of model performance: accuracy and the variability among replicate models fits. Accuracy was assessed by the experts (see question 3 above). This can be considered a measure of discrimination ability because the experts based their judgements on whether habitat suitability was predicted to be higher at more suitable locations and vice versa. The variability among replicate model fits was calculated as the sum of the variance of habitat suitability across grid cells (hereafter "variance"). This measure includes the variability among algorithms and models fitted to different data subsets. Hence, it is in part a measure of sampling variability—i.e. the variation in some statistic, here habitat suitability—among samples.

Explanatory variables

We assumed that SDM accuracy and variance are functions of five variables: species' prevalence (see below), sample size, expert-assessed niche completeness (the degree to which the available records cover the species' environmental niche), expert-assessed range completeness (as niche completeness but for the species' geographic range) and expert assessor identity. There are many variables that have been shown to affect SDM accuracy—such as the choice of SDM algorithm, covariates and pseudo-absence strategy—that we do not consider here. We hold these variables constant in our SDM workflow so have effectively conditioned on them.

The explanatory variables were measured or derived in different ways. Range and niche completeness were assessed by the experts and reported on a five-point Likert scale as described above. Sample size is simply the number of 1 km grid cells (EPSG:27700) in which each species was recorded from 2000–2015 (i.e. an imperfect measure of range size). We use the term prevalence to describe an index of species' range size that corrects for survey incompleteness, based on expert-assessed range completeness. Specifically, prevalence equals sample size divided by range completeness. Prevalence is low where sample size is low and expert-judged range completeness is

high (i.e. where recorded range size at 1km is low despite the fact that a high proportion of the species' range has been sampled), and vice versa.

Conceptual models

We used DAGs to conceive plausible conceptual models depicting the effects of the explanatory variables on SDM accuracy and variance. DAGs are non-parametric, and are distinct from the statistical models used to analyse them (see "Statistical analysis of conceptual models" below). Our general strategy was to start with a theoretically plausible DAG, test whether it was empirically plausible, then refine it accordingly (similar to steps 1-3 in Grace & Irvine, 2020). The primary goal of model testing was to ascertain whether a DAG's (conditional) independencies were consistent with our data. If these were consistent, we then assessed the support for the DAG's implied mediation pathways using the "joint significance" method (MacKinnon et al., 2002). At no point did we posit a theoretically implausible DAG just to satisfy these criteria.

Using this strategy, we rejected two DAGs then arrived at two DAGs that were both theoretically and empirically plausible. It was not clear which of these models was the most plausible, so we consider them both hereafter. Full details of the model conceptualisation and testing process can be found in the R Markdown document in supplementary material two. Justifications for the DAG structures are given below.

The first of the plausible DAGs, hereafter "Model one" (Fig. 2), supposes that SDM accuracy is caused by all five of the explanatory variables described above. Sample size is assumed to have a direct effect; this effect has been reported across a range of species with varying characteristics (Wisz et al., 2008). Sample size itself is assumed to be caused by species' prevalence and range completeness: for a given prevalence, larger range completeness equals higher sample size; likewise, for a given range completeness, larger prevalence equals larger sample size. Hence, prevalence and range completeness have indirect effects on accuracy mediated by sample size. Prevalence is assumed to have a direct effect on accuracy, as reported by Santika (2011). Prevalence also has direct effects on range completeness and niche completeness: it is more difficult for a recorders to cover a given proportion of a prevalent species' range and niche than a common one. Hence, prevalence has an indirect on accuracy mediated by range and niche completeness, and, because range completeness has an effect on sample size, by sample size. Niche completeness has a direct effect on accuracy: SDMs estimate species' environmental niches, so it is logical to assume that sufficient coverage of those niches will affect their ability to do so. Niche completeness and range completeness are assumed to have a common cause in "recorder behaviour", which is defined as recorders' decisions about where to sample geographically and hence environmentally. The recorder behaviours determining this pattern in our aggregated datasets are unobserved.

In terms of variance, Model one supposes that sample size has a direct effect. Recall that we calculated variance across replicate model fits, which varied in terms of algorithm and occurrence data. Grimmett et al. (2020) showed that the variability among SDM predictions from different SDM algorithms was lower at high sample sizes. Likewise, variability among models fitted to data subsets should be lower at high sample sizes. This is because smaller samples are more likely to be unusual (different from the population) by chance, which increases the sampling variability (Lohr, 2022).

In addition to sample size, model one assumes that species' prevalence has a direct effect on SDM variance. Syphard & Franklin (2009) showed that species' prevalence affects the congruence of spatial predictions among SDM algorithms, which is one component of our measure of SDM variance. Furthermore, we suspect that variability among model fits using the same algorithm will be

larger for prevalent species. Our theoretical justification for this effect is that there is less variability in the types of habitats in which rarer species occur; no matter which subsample of the data is considered, there will only be occurrence data from those habitats. Hence, the models will produce more consistent predictions (when sample size is conditioned on).

Niche completeness is the final variable assumed to have a direct effect on variance by Model one. We suspect that different SDM algorithms will respond differently to low niche completeness, which will result in increased variance among the predictions from those algorithms. For example, random forests and Maxent can be relatively complex, so are able to fit—or overfit—the available data closely. On the other hand, the regularized GLMs do not overfit the data by definition. This means that, where niche completeness is low, Maxent and random forests can be expected to produce different predictions to the GLMs in geographic locations and environments that have not been sampled (Werkowska et al., 2017), which would be reflected by an increase in our measure of variance.

After preliminary testing (supplementary material two), the data provided mixed support for the role of sample size in Model one. There was evidence that sample size mediated the effect of prevalence on accuracy. However, there was only weak evidence for an effect of range completeness mediated through sample size (supplementary material 2, p. 12), and models including sample size did not explain appreciably more of the variance in accuracy than those without it (see "Results" and supplementary material 2, p. 21). Hence, we devised a model (hereafter "Model two") in which it is assumed that sample size has no effect on accuracy (Fig. 3). That is to say, Model two supposes that range completeness has only a direct effect on accuracy, and none of the effect of prevalence on accuracy is mediated through sample size.

Statistical analysis of conceptual models

We used piecewise path analysis to estimate the effects of the explanatory variables described above on SDM accuracy and variance, using the R package *piecewiseSEM* (Lefcheck, 2016). Path analysis is the process of estimating path coefficients for each arrow, or "edge", in a DAG (Grace, 2006). They are equivalent to the coefficients estimated by regressing the variable on the receiving end of an edge on the variable from which the edge originates; that is to say, by regressing the "child" on its "parent" in DAG parlance. Where one variable affects another via more than one pathway (i.e. where a child has more than one parent), the path coefficient for one parent is equal to the partial regression coefficient obtained by regressing the child on that parent whilst conditioning on all other parents (i.e. multiple regression). In our analysis, for ease of interpretation, we standardised the path coefficients using the z transformation.

Path coefficients indicate the direct effect of each parent on its child, but these can be used to calculate indirect and total effects (Sobel, 1982). One variable has an indirect effect on another where there is an intermediate variable (mediator). Indirect effects may be subdivided into specific and total indirect effects. A specific indirect effect is the product of all path coefficients in one pathway; for example, prevalence → n → accuracy in Fig. 1. The total indirect effect of one variable on another is the sum of the specific indirect effects over all pathways linking the them (Preacher & Hayes, 2008; Tarling, 2009). The total effect of one variable on another is the sum of its direct effect and total indirect effect (Grace, 2006; Tarling, 2009).

To assess the uncertainty associated with the estimated effects, we used nonparametric bootstrapping. We resampled the data with replacement to create 1000 bootstrap samples, fitted models to each sample, and report the 95% (percentile) confidence intervals for each effect across samples.

One might expect the expert-assessed variables in our analysis to differ systematically among taxon groups and assessors (recalling that one assessor evaluated the models for each taxon group). For example, the experts might simply differ in what they perceive to be an accurate model, or what constitutes "very good" coverage of a species' range. Or perhaps expert-assessed accuracy will vary between taxon groups if, say, the environmental covariates are more appropriate for some groups than others.

To assess the extent of any systematic differences between taxon groups in terms of expert-scored accuracy, we calculated their intraclass correlation coefficients. The respective values 0.08, 0.25 and 0.23 (supplementary material 2, p. 21), indicating that the data are not independent within assessors. Hence, we include a random intercept for assessor identity in the portions of Models one and two in which accuracy, range completeness or niche completeness are the response variable.

Sensitivity analysis

Piecewise path models are based on linear regression and so are bound by the same assumptions. These include the assumptions that the response variables are numeric and normally distributed, which our data violate. Nevertheless, we proceeded with piecewise path models because it has been oft demonstrated that linear regression is robust to such violations (e.g. Norman, 2010).

The robustness of linear regression notwithstanding, we assessed the sensitivities of our results to the choice of analytical method. By analytical method we mean statistical model, which is different to the non-parametric causal models described above. We analysed both causal models using several analytical methods, which varied in terms of how they treat the response variable expert score (ordinal or numeric), how they accounted for assessor identity (either by complete pooling, random intercepts or fixed effects), and how model fitting was achieved (e.g. covariance- or piecewise least-squares-based). Four of the five additional analytical methods gave roughly identical results (supplementary material three), so we only present the results from the multilevel piecewise path models here.

Results

Models one and two are highly congruent. They explain identical proportions (to two decimal places) of the variation in accuracy (0.16) and variance (0.35). The models also generally agree on the directions and magnitudes of the effects of each explanatory variable on accuracy and variance, and whether the confidence intervals span zero. The only exception is the effect of prevalence on accuracy. The standardised path coefficient is -0.23 in model one, and -0.05 in model two. Both coefficients include zero in their 95% confidence intervals, however.

Both models agree that niche completeness is the only explanatory variable to have a direct effect on accuracy (Table 2). This effect is positive, as one would expect from ecological theory. Both models also suggest that prevalence has a total indirect effect on SDM accuracy (Table 2). We present this effect for transparency, but it cannot be meaningfully interpreted. We expand on this point in the Discussion below.

Models one and two make the same assumptions about what determines variance, so their path coefficients are identical. Sample size, species' prevalence and niche completeness all have strong direct effects on variance. The effect of niche completeness is negative—as it increase variance decreases. The effects of species' prevalence and sample size are positive—as they increases variance does too. That is, predictions for widespread species with large sample sizes are most variable. Prevalence has the strongest direct effect. Prevalence and range completeness also have indirect effects on variance; prevalence has the stronger of these effects.

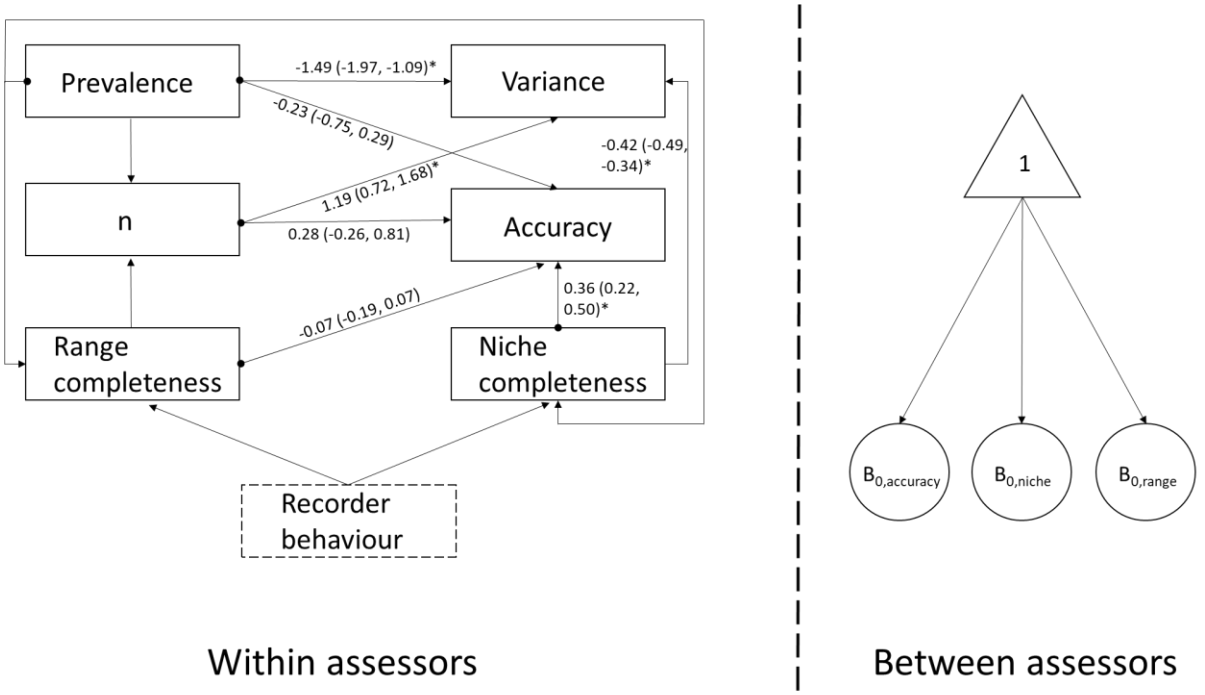

Figure 2. Directed Acyclic Graph (Greenland et al., 1999) depicting Model one's assumptions about what determines SDM accuracy and precision. The right-hand portion of the figure indicates that we included random intercepts in the model. Specifically, we allowed the intercepts to vary by assessor identity (and hence taxon group) in the portions of the model in which range completeness, niche completeness or accuracy is the response variable, because these were all expert assessed. The path coefficients were estimated using piecewise path analysis and were standardised using the z transformation. Bootstrapped 95% confidence intervals are shown in the parentheses; asterisks indicate that these do not span zero. For ease of interpretation, we only present the path coefficients for edges leading to accuracy and variance, because these are the variables of interest.

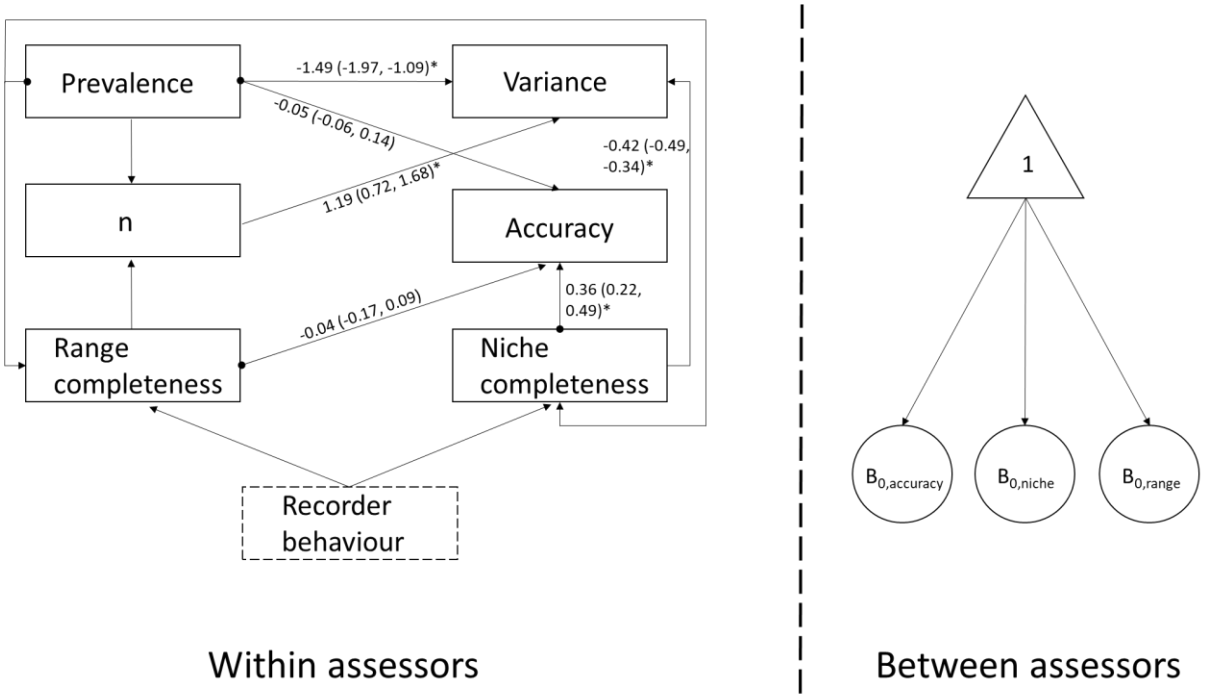

Figure 3. Directed Acyclic Graph (Greenland et al., 1999) depicting Model two's assumptions about what determines SDM accuracy and precision. The right-hand portion of the figure indicates that we included random intercepts in the model. Specifically, we allowed the intercepts to vary by assessor identity (and hence taxon group) in the portions of the model in which range completeness, niche completeness or accuracy is the response variable, because these were all expert assessed. The path coefficients were estimated using piecewise path analysis and were standardised using the z transformation. Bootstrapped 95% confidence intervals are shown in the parentheses; asterisks indicate that these do not span zero. For ease of interpretation, we only present the path coefficients for edges leading to accuracy and variance, because these are the variables of interest.

Table 2. Direct, indirect and total effects of the explanatory variables on the response variables expert score and uncertainty from models one and two. Total indirect effects are the sum of the specific direct effects in each pathway linking one variable to another; the specific direct effects for each pathway are the product of the path coefficients in that pathway. Total effects are the sum of the direct and total indirect effects. 95% confidence intervals are given in parentheses. Confidence intervals were obtained by bootstrapping, but the R^2 values and effect point estimates are from the models fitted to the original data. Red-filled cells indicate effects that are likely to be confounded (see the "Discussion" for more on this).

Model	Response	R^2	Explanatory variable	Direct effect	Total indirect effect	Total effect
One	Accuracy	0.16	n	0.26 (-0.21, 0.78)	-	0.26 (-0.83, 0.37)
			Niche completeness	0.36 (0.22, 0.50)	-	0.36 (0.23, 0.50)
			Prevalence	-0.23 (-0.75, 0.29)	0.34 (-0.19, 0.86)	0.11 (0.01, 0.20)
			Range completeness	-0.07 (-0.19, 0.07)	0.02 (-0.02, 0.05)	-0.05 (-0.17, 0.08)
	Variance	0.35	n	1.19 (0.72, 1.68)	-	1.19 (0.72, 1.68)
			Niche completeness	-0.42 (-0.49, -0.34)	-	-0.42 (-0.49, -0.34)
			Prevalence	-1.49 (-1.95, -1.09)	1.10 (0.63, 1.56)	-0.41 (-0.46, -0.36)
			Range completeness	-	0.08 (0.05, 0.11)	0.08 (0.05, 0.11)
Two	Accuracy	0.16	Niche completeness	0.36 (0.22, 0.50)	-	0.36 (0.22, 0.50)
			Prevalence	0.05 (-0.05, 0.14)	0.06 (0.10, 0.21)	0.11 (0.01, 0.21)
			Range completeness	-0.04 (-0.17, 0.08)	-	-0.04 (-0.17, 0.08)
	Variance	0.35	n	1.19 (0.72, 1.68)	-	1.19 (0.72, 1.68)
			Niche completeness	-0.42 (-0.49, -0.34)	-	-0.42 (-0.49, -0.34)
			Prevalence	-1.49 (-1.95, -1.09)	1.10 (0.63, 1.56)	-0.41 (-0.46, -0.36)

			Range completeness	-	0.08 (0.05, 0.11)	0.08 (0.05, 0.11)

Discussion

In this paper, we used expert validation, graph theory and causal analysis to shed light on the drivers of SDM performance. We considered two components of model performance: accuracy, as assessed by the experts; and the variance among replicate model fits. We constructed DAGs depicting the effects of various explanatory variables on SDM performance, then analysed those DAGS using piecewise path models.

We suggest that the experts' knowledge is likely to be more informative than any one dataset that could have been used for model validation. Each expert is a national curator of the data for their taxon group. As such, they have much local, national and international field knowledge, and have written about the focal species' autecologies (e.g. for distribution atlases and in field guides). Hence, their assessments arguably reflect an unrivalled synthesis of information.

Our models suggest that prevalence has total indirect effects on SDM accuracy and variance, but these cannot be meaningfully interpreted. The problem is that recorder behaviour is unobserved, and it is difficult to see how it could be estimated (it covers e.g. geographic sampling biases and preferential sampling of some species in some locations). This means that the effects of prevalence on range completeness and niche completeness are confounded because it was not possible to condition on recorder behaviour. In turn, this is likely to have biased the total indirect effects of prevalence on accuracy and variance. That said, the structure of the DAGs is such that this will not bias any other effects reported; on the contrary, assuming our DAGs are true, inclusion of range and niche completeness enables unbiased estimation of the remainder of the path coefficients.

Putting the above to one side, we found that niche completeness was the only explanatory variable to have an effect on expert-assessed accuracy (Table 2). Ours is not the first study to report this effect. For example, Konowalik & Nosol (2021) showed that SDMs fitted to datasets with greater environmental coverage generally produced models with greater discrimination abilities for one species of plant, as indicated by AUC and an expert assessor. However, we have demonstrated this on a much larger scale (i.e. across 534 species) and in an explicitly causal framework.

Contrary to previous studies, we found little evidence that sample size affects SDM accuracy (Fig. 3). Model one includes an effect of sample size on accuracy, whereas Model two does not; nevertheless, Model one does not explain more of the variance in accuracy (Table 2). Furthermore, in model one the 95% confidence interval for the path coefficient denoting the effect of sample size on accuracy spans zero. We analysed Model one using five additional analytical methods (e.g. cumulative link ordinal regression, covariance-based path models) in supplementary material 3. Under the admittedly arbitrary assumption that $p > 0.2$ indicates little evidence for an effect, four of the five additional methods concur that there is little evidence for an effect of sample size on SDM accuracy, given the assumptions of Model one.

We suggest that the previously reported associations between sample size and SDM accuracy are likely to reflect improved coverage of species' environmental niches at higher sample sizes; that is to say, in causal terms, sample size is a confounder of the direct effect of niche coverage on SDM accuracy. For example, Wisz et al. (2008) and Feeley & Silman (2011) subsampled datasets for several species and showed that models fitted to the smaller subsets were less accurate. In both cases, however, the authors acknowledged that coverage of species' environmental niches was lower in the smaller samples, clearly demonstrating the potential for confounding. However, neither

paper discussed this in any detail, instead focusing on sample size without reference to niche completeness. We suspect that the same is true of other studies, particularly those which did not disclose variation in niche completeness with sample size (e.g. Hernandez et al., 2006).

Another study to have claimed an effect of sample size on SDM accuracy is the seminal paper of Stockwell & Peterson (2002). Like the studies mentioned above, these authors subsampled complete datasets to manipulate sample size. However, they then created training datasets with n = 1000 presences by resampling these subsamples with replacement. Hence, they actually held sample size constant, but varied the degree to which those samples reflected the full datasets. It is likely that samples more closely resembling the full dataset covered more of each species' environmental niche, again demonstrating the potential for confounding.

The spurious effect of sample size on SDM accuracy can be demonstrated using our data (final section in supplementary material three). We regressed SDM accuracy on sample size, and the coefficient was highly significant. We then included niche completeness as an additional independent variable, and the effect of sample size became highly insignificant and reduced in magnitude. This result was evident using both ordinary linear regression, and cumulative link ordinal regression with random intercepts for assessor ID.

Other than the confounding effect of niche completeness, there are two alternative explanations for the discrepancy between ours and others' findings about the effects of sample size on SDM accuracy. The first is that we did not fit SDMs for species recorded in fewer than ten grid cells. We took this decision because we fitted the models using five-fold cross validation, which further reduced the sample size for any given fit. It is possible, however, that SDM accuracy is more sensitive to sample size where $n < 10$.

Another explanation for the lack of an effect of sample size on accuracy could be that we omitted an important explanatory variable in Model one. An obvious example is niche breadth. It has been reported that niche breadth is negatively associated with SDM accuracy (Tessarolo et al., 2021), and one might reasonably assume that it is positively associated with sample size. Using the rules of omitted variable bias (Angrist & Pischke, 2009), it can be shown that in these circumstances our estimate for the effect of sample size on accuracy would be negatively biased.

Whilst omitting niche breadth from Model one could negatively bias the estimated effect of sample size on SDM accuracy, we suspect that the extent of this bias would be small. We calculated the number of land cover classes (Morton et al., 2011) on which each species was recorded as a proxy for its niche breadth. This is not a perfect proxy for niche breadth, particularly for the invertebrates, but we suspect that it is a reasonably strong correlate thereof at the scale of our models (1 km^2). The number of land cover classes on which species have been recorded is very weakly correlated with the residuals from Model one ($r = 0.08$, $p = 0.07$; supplementary material three), which implies that the extent of the omitted variable bias is small. This could reflect the fact that niche breadth is likely to be strongly correlated with species' prevalence (Slatyer et al., 2013), which we do include in the model. Hence, prevalence should explain a similar portion of the variance in accuracy to niche breadth.

Alternative explanations notwithstanding, we suggest that the confounding effect of niche completeness is the most logical explanation for our finding that sample size has little effect on SDM accuracy. This is worrying because analysts frequently use sample size as the sole criterion when deciding whether or not to fit SDMs for a given species (e.g. Amini Tehrani et al., 2021; Hoveka et al., 2020, 2022; Spiers et al., 2018; Zellmer et al., 2019). We agree with Santini et al., (2021), who noted

that, of the studies making methodological recommendations in the SDM literature, those making convenient recommendations (e.g. proceed if you have a sample size of at least *n*) tend to be more favourably received and widely cited. We appeal to analysts to think more critically and consider more nuanced (and ecological!) aspects of their data such as niche completeness.

Assessing niche completeness is more difficult than calculating sample size, but there are several ways that one might go about this. One option is to consult experts as we did here. Another is to use range completeness as a proxy for niche completeness on the assumption that these are highly correlated; the analyst could then compare the distribution of records to published range maps, for example. Tools to assess the environmental representativeness of species occurrence data also exist (e.g. Boyd et al., 2021). Where additional data thought to cover a species' niche are available—e.g. coarse-scale data from an atlas, or a digitised range map—these tools could be used to calculate niche coverage relative to the more complete data.

Whilst we found little evidence for an effect of sample size on SDM accuracy, we found that SDM variance increased with sample size (when controlling for the other covariates; the raw correlation was negative). It seems unlikely that increasing sample size results in greater sampling variability in estimated habitat suitability scores. Rather, the effect of sample size on our measure of variance probably reflects increased inter-algorithm variability at high sample sizes (but see Grimmett et al. 2020).

One explanation for increased inter-algorithm variability at large sample sizes could be an increased disparity in algorithm complexity. For example, as sample size increases, Maxent will consider an increasing number of "feature classes" (Merow et al., 2013), which are essentially response curve shapes. This increase in complexity could result in predictions that differ from the simpler GLMs, thus explaining the increase in our measure of variance with sample size.

In addition to sample size, we also found that species' prevalence has a strong effect on SDM variance: models for widespread species tend to be less precise. This is a feature of the species' autecology and not something that the analyst can change. Again, then, we suggest that priority should be given to collating data covering as much of the focal species' environmental niche as possible, thereby increasing the chance that the model will be accurate. Indeed, this will also increase the chance that the SDM is precise (Table 2).

An important implication of our results is that the common practice of "stacking" individual species' SDMs to estimate species richness or similar is a risky business. Model performance is not random; rather, as we have shown, it varies with species traits and data characteristics. Hence, there is no reason to suppose that the errors will average out over many species. There could be serious biases in play.

We do not claim that our causal models are true. However, in depicting them as DAGs we have laid bare our assumptions about what determines SDM performance in a falsifiable manner. We believe that this is an improvement on much of the (vast) literature proffering advice on fitting SDMs, and that it clarifies the causal basis of much of this advice in a way that can be built upon clearly.

Acknowledgements

R.J.B., M.H, D.B.R and O.L.P. were supported by the NERC award number NE/R016429/1 as part of the UK Status, Change and Projections of the Environment (UK- SCAPE) program delivering National Capability.

Data availability

The ensemble habitat suitability surfaces are embargoed until March 4th 2023, at which point they will become available at https://doi.org/10.5285/ec921bc2-5538-47ed-9e72-0d687b4ca4d3. We will provide the expert scores for these models when this article is accepted for publication.

Author contributions

R.J.B.: Conceptualization (equal), Methodology (lead), Formal analysis (lead), Writing – original draft (lead), Writing – review & editing (lead), Visualization (lead).

M.H.: Investigation (equal); Writing – review & editing (supporting); Data curation (equal).

D.B.R.: Project administration (lead); Writing – review & editing (supporting).

T.B.: Investigation (equal)

K.A.H.: Investigation (equal); Writing – review & editing (supporting).

C.R.M: Investigation (equal); Writing – review & editing (supporting).

R.K.A.M.: Investigation (equal); Writing – review & editing (supporting).

C.P: Investigation (equal)

S.P: Investigation (equal)

C.D.P: Investigation (equal); Writing – review & editing (supporting).

P.T.: Investigation (equal)

R.W: Investigation (equal); Writing – review & editing (supporting).

S.G.B.: Investigation (equal)

O.L.P.: Conceptualization (equal), Methodology (supporting), Formal analysis (lead), Writing – original draft (supporting), Writing – review & editing (supporting); Data curation (equal).

References

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics : An Empiricist ' s Companion* (Issue March). Princeton Univ. Press.

Arenas-Castro, S., Regos, A., Martins, I., Honrado, J., & Alonso, J. (2022). Effects of input data sources on species distribution model predictions across species with different distributional ranges. *Journal of Biogeography*, *49*(7), 1299–1312. https://doi.org/10.1111/jbi.14382

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, *3*(2), 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

Baron, R., & Kenny, D. (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182. https://doi.org/10.1037//0022-3514.51.6.1173

Bazzichetto, M., Lenoir, J., Da Re, D., Tordoni, E., Rocchini, D., Malavasi, M., Vojtech, B., & Sperandii, M. G. (2022). Effect of sampling strategies on the response curves estimated by plant species

distribution models. *Ecoevorxiv*.

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, *19*, 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002

Boyd, R., Pescott, O., Ball, S., Barber, T., Boardman, P., Fox, R., Harrower, C., Harvey, M., Haysom, K., Julian, A., MacAdam, C., Mathews, F., Morris, R., Palmer, S., Preston, C., Roy, D., Simkin, J., Taylor, P., Walker, K., & Ward, R. (2022). *UK maps of habitat suitability surfaces at 1km resolution for mammals, lichens, bryophytes, plants and invertebrates 2000-2015.* NERC EDS Environmental Information Data Centre. https://doi.org/https://doi.org/10.5285/ec921bc2-5538-47ed-9e72-0d687b4ca4d3

Boyd, Robin, Powney, G., Carvell, C., & Pescott, O. L. (2021). occAssess: An R package for assessing potential biases in species occurrence data. *Ecology and Evolution*, *September*. https://doi.org/10.1002/ece3.8299

Breiman, T., Cutler, A., & Classification, D. (2018). *Package ' randomForest .'* https://doi.org/10.1023/A

Bucklin, D. N., Basille, M., Benscoter, A. M., Brandt, L. A., Mazzotti, F. J., Romañach, S. S., Speroterra, C., & Watling, J. I. (2015). Comparing species distribution models constructed with different subsets of environmental predictors. *Diversity and Distributions*, *21*(1), 23–35. https://doi.org/10.1111/ddi.12247

Cerasoli, F., Iannella, M., D'Alessandro, P., & Biondi, M. (2017). Comparing pseudo-absences generation techniques in Boosted Regression Trees models for conservation purposes: A case study on amphibians in a protected area. *PLoS ONE*, *12*(11), 1–23. https://doi.org/10.1371/journal.pone.0187589

Chapman, D., Pescott, O. L., Roy, H. E., & Tanner, R. (2019). Improving species distribution models for invasive non-native species with biologically informed pseudo-absence selection. *Journal of Biogeography*, *46*(5), 1029–1040. https://doi.org/10.1111/jbi.13555

De Marco, P., & Nóbrega, C. C. (2018). Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation. *PLoS ONE*, *13*(9). https://doi.org/10.1371/journal.pone.0202403

Dudík, M., Schapire, R. E., & Phillips, S. J. (2005). Correcting sample selection bias in maximum entropy density estimation. *Advances in Neural Information Processing Systems*, 323–330.

El-Gabbas, A., Gilbert, F., & Dormann, C. F. (2020). Spatial conservation prioritisation in data-poor countries: A quantitative sensitivity analysis using multiple taxa. *BMC Ecology*, *20*(1), 1–12. https://doi.org/10.1186/s12898-020-00305-7

Feeley, K. J., & Silman, M. R. (2011). Keep collecting: Accurate species distribution modelling requires more collections than previously thought. *Diversity and Distributions*, *17*(6), 1132–1140. https://doi.org/10.1111/j.1472-4642.2011.00813.x

Fourcade, Y., Engler, J., Ro¨dder, D., & 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*(5), 1–13. https://doi.org/10.1371/journal.pone.0097122

Fukuda, S., & De Baets, B. (2016). Data prevalence matters when assessing species' responses using data-driven species distribution models. *Ecological Informatics*, *32*, 69–78. https://doi.org/10.1016/j.ecoinf.2016.01.005

584 Grace, J. B. (2006). *Structural equation modeling and natural systems*. Cambridge University Press.

585 Grace, J. B., & Irvine, K. M. (2020). Scientist's guide to developing explanatory statistical models
586 using causal analysis principles. *Ecology*, *101*(4), 1–14. https://doi.org/10.1002/ecy.2962

587 Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. In
588 *Epidemiology* (Vol. 10, Issue 1, pp. 37–48). https://doi.org/10.1097/00001648-199901000-
589 00008

590 Hao, T., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2020). Testing whether ensemble
591 modelling is advantageous for maximising predictive performance of species distribution
592 models. *Ecography*, *43*(4), 549–558. https://doi.org/10.1111/ecog.04890

593 Hernandez, P. A., Graham, C. H., Master, L. L., & Albert, D. L. (2006). The effect of sample size and
594 species characteristics on performance of different species distribution modeling methods.
595 *Ecography*, *29*(5), 773–785. https://doi.org/10.1111/j.0906-7590.2006.04700.x

596 Hijmans, R. J., Phillips, S. J., Leathwick, J. R., & Elith, J. (2017). *dismo: Species Distribution Modeling. R*
597 *package version 1.1-4.* https://cran.r-project.org/package=dismo

598 Jiménez-valverde, A., Acevedo, P., Barbosa, A. M., Lobo, J. M., & Real, R. (2013). Discrimination
599 capacity in species distribution models depends on the representativeness of the
600 environmental domain. *Global Ecology and Biogeography*, *22*, 508–516.
601 https://doi.org/10.1111/geb.12007

602 Konowalik, K., & Nosol, A. (2021). Evaluation metrics and validation of presence-only species
603 distribution models based on distributional maps with varying coverage. *Scientific Reports*,
604 *11*(1), 1–15. https://doi.org/10.1038/s41598-020-80062-1

605 Leroy, B., Delsol, R., Hugueny, B., Meynard, C. N., Barhoumi, C., Barbet-Massin, M., & Bellard, C.
606 (2018). Without quality presence–absence data, discrimination metrics such as TSS can be
607 misleading measures of model performance. *Journal of Biogeography*, *45*(9), 1994–2002.
608 https://doi.org/10.1111/jbi.13402

609 Lobo, J. M., Jiménez-Valverde, A., & Hortal, J. (2010). The uncertain nature of absences and their
610 importance in species distribution modelling. *Ecography*, *33*(1), 103–114.
611 https://doi.org/10.1111/j.1600-0587.2009.06039.x

612 Lobo, J. M., Jiménez-valverde, A., & Real, R. (2008). AUC: A misleading measure of the performance
613 of predictive distribution models. *Global Ecology and Biogeography*, *17*(2), 145–151.
614 https://doi.org/10.1111/j.1466-8238.2007.00358.x

615 MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of
616 methods to test mediation and other intervening variable effects. *Psychological Methods*, *7*(1),
617 83–104. https://doi.org/10.1037/1082-989X.7.1.83

618 Merow, C., Smith, M. J., & Silander, J. A. (2013). A practical guide to MaxEnt for modeling species'
619 distributions: What it does, and why inputs and settings matter. *Ecography*, *36*(10), 1058–1069.
620 https://doi.org/10.1111/j.1600-0587.2013.07872.x

621 Montalva, J., Sepulveda, V., Vivallo, F., & Silva, D. P. (2017). New records of an invasive bumble bee
622 in northern Chile: expansion of its range or new introduction events? *Journal of Insect*
623 *Conservation*, *21*(4), 657–666. https://doi.org/10.1007/s10841-017-0008-x

624 Morton, R. D., Rowland, C., Wood, C., Meek, L., Marston, G., Smith, G., Wadsworth, R., & Simpson, I.
625 (2011). *Land Cover Map 2007 (1km percentage target class, N. Ireland).*
626 https://doi.org/https://doi.org/10.5285/e611794a-2f7c-4cfc-a8ab-4c38131e0fad

Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Sciences Education*, *15*(5), 625–632. https://doi.org/10.1007/s10459-010-9222-y

Outhwaite, C., Powney, G., August, T., Chandler, R., Rorke, S., Pescott, O. L., Harvey, M., Roy, H. E., Fox, R., Roy, D. B., Alexander, K., Ball, S., Bantock, T., Barber, T., Beckmann, B. C., Cook, T., Flanagan, J., Fowles, A., Hammond, P., … Isaac, N. J. B. (2019). Annual estimates of occupancy for bryophytes, lichens and invertebrates in the UK, 1970-2015. *Scientific Data*, *6*(1), 259. https://doi.org/10.1038/s41597-019-0269-1

Pearl, J., Glymour, M., & Jewell, N. (2016). *Causal inference in statistics: A primer*. Wiley.

Pescott, O. L. (2022). *A Google Sheets-linked R Shiny app for the expert validation of Species Distribution Models (Version 1)*. Zenodo. https://doi.org/https://doi.org/10.5281/zenodo.7082588

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

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*(3), 879–891. https://doi.org/10.3758/BRM.40.3.879

Santika, T. (2011). Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. *Global Ecology and Biogeography*, *20*(1), 181–192. https://doi.org/10.1111/j.1466-8238.2010.00581.x

Simons, R. R. L., Croft, S., Rees, E., Tearne, O., Arnold, M. E., & Johnson, N. (2019). *Using species distribution models to predict potential hot-spots for Rift Valley Fever establishment in the United Kingdom*.

Slatyer, R. A., Hirst, M., & Sexton, J. P. (2013). Niche breadth predicts geographical range size: A general ecological pattern. *Ecology Letters*, *16*(8), 1104–1114. https://doi.org/10.1111/ele.12140

Smart, S. M., Jarvis, S. G., Mizunuma, T., Herrero-Jáuregui, C., Fang, Z., Butler, A., Alison, J., Wilson, M., & Marrs, R. H. (2019). Assessment of a large number of empirical plant species niche models by elicitation of knowledge from two national experts. *Ecology and Evolution*, *9*(22), 12858–12868. https://doi.org/10.1002/ece3.5766

Sobel, M. E. (1982). *Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models*. *13*(1982), 290–312.

Steen, V. A., Elphick, C. S., & Tingley, M. W. (2019). An evaluation of stringent filtering to improve species distribution models from citizen science data. *Diversity and Distributions*, *25*(12), 1857–1869. https://doi.org/10.1111/ddi.12985

Steen, V. A., Tingley, M. W., Paton, P., & Elphick, C. (2020). Spatial thinning and class balancing : Key choices lead to variation in the performance of species distribution models with citizen science data. *Methods in Ecology and Evolution*, *December*. https://doi.org/10.1111/2041-210X.13525

Stewart, S. B., Fedrigo, M., Kasel, S., Roxburgh, S. H., Choden, K., Tenzin, K., Allen, K., & Nitschke, C. R. (2022). Predicting plant species distributions using climate-based model ensembles with corresponding measures of congruence and uncertainty. *Diversity and Distributions*, *28*(5), 1105–1122. https://doi.org/10.1111/ddi.13515

Stockwell, D. R. B., & Peterson, A. T. (2002). Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, *148*(1), 1–13. https://doi.org/10.1016/S0304-3800(01)00388-X

Tarling, R. (2009). Statistical Modelling for Social Researchers: Principles and practices. In *Canadian Journal of Sociology* (Vol. 34, Issue 1). Taylor & Francis Group. https://doi.org/10.29173/cjs4634

Tessarolo, G., Lobo, J. M., Rangel, T. F., & Hortal, J. (2021). High uncertainty in the effects of data characteristics on the performance of species distribution models. *Ecological Indicators*, *121*, 107147. https://doi.org/10.1016/j.ecolind.2020.107147

Watling, J. I., Brandt, L. A., Bucklin, D. N., Fujisaki, I., Mazzotti, F. J., Romañach, S. S., & Speroterra, C. (2015). Performance metrics and variance partitioning reveal sources of uncertainty in species distribution models. *Ecological Modelling*, *309–310*(August), 48–59. https://doi.org/10.1016/j.ecolmodel.2015.03.017

Werkowska, W., Márquez, A. L., Real, R., & Acevedo, P. (2017). A practical overview of transferability in species distribution modeling. *Environmental Reviews*, *25*(1), 127–133. https://doi.org/10.1139/er-2016-0045

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., Elith, J., Dudík, M., Ferrier, S., Huettmann, F., Leathwick, J. R., Lehmann, A., Lohmann, L., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M. C., … Zimmermann, N. E. (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, *14*(5), 763–773. https://doi.org/10.1111/j.1472-4642.2008.00482.x

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