


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

Robin J. Boyd¹  | Martin Harvey¹ | David B. Roy¹ | Tony Barber² | Karen A. Haysom³ | Craig R. Macadam^{4,5,6} | Roger K. A. Morris^{4,7} | Carolyn Palmer⁸ | Stephen Palmer⁸ | Chris D. Preston⁹ | Pam Taylor¹⁰ | Robert Ward³ | Stuart G. Ball¹¹ | Oliver L. Pescott¹

¹UK Centre for Ecology and Hydrology, Crowmarsh Gifford, UK

²British Myriapod and Isopod Group, Ipswich, UK

³Amphibian and Reptile Conservation, Bournemouth, UK

⁴The Natural History Museum, London, UK

⁵Riverfly Recording Schemes, Stirling, UK

⁶Unit 4, Beta Centre, Stirling University Innovation Park, Stirling, UK

⁷Hoverfly Recording Scheme, Mitcham, UK

⁸Gelechiid Recording Scheme, Preston, UK

⁹Cambridge, UK

¹⁰British Dragonfly Society, Huntingdon, UK

¹¹Hoverfly Recording Scheme, Peterborough, UK

Correspondence

Robin J. Boyd, UK Centre for Ecology and Hydrology, Benson Ln, Crowmarsh Gifford, Oxfordshire, UK.
Email: robboy@ceh.ac.uk

Funding information

Natural Environment Research Council, Grant/Award Number: NE/R016429/1

Editor: Luigi Maiorano

Abstract

Aim: To develop a causal understanding of the drivers of Species distribution model (SDM) performance.

Location: United Kingdom (UK).

Methods: We measured the accuracy and variance of SDMs fitted for 518 species of invertebrate and plant in the UK. Our measure of variance reflects variation among replicate model fits, and taxon experts assessed model accuracy. Using directed acyclic graphs, we developed a causal model depicting plausible effects of explanatory variables (e.g. species' prevalence, sample size) on SDM accuracy and variance and quantified those effects using a multilevel piecewise path model.

Results: According to our model, sample size and niche completeness (proportion of a species' niche covered by sampling) directly affect SDM accuracy and variance. Prevalence and range completeness have indirect effects mediated by sample size. Challenging conventional wisdom, we found that the effect of prevalence on SDM accuracy is positive. This reflects the facts that sample size has a positive effect on accuracy and larger sample sizes are possible for widespread species. It is possible, however, that the omission of an unobserved confounder biased this effect. Previous studies, which reported negative correlations between prevalence and SDM accuracy, conditioned on sample size.

Main conclusions: Our model explicates the causal basis of previously reported correlations between SDM performance and species/data characteristics. It also suggests that niche completeness has similarly large effects on SDM accuracy and variance as sample size. Analysts should consider niche completeness, or proxies thereof, in addition to sample size when deciding whether modelling is worthwhile.

KEYWORDS

causal inference, directed acyclic graph, expert elicitation, species distribution modelling, structural equation modelling

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Diversity and Distributions* published by John Wiley & Sons Ltd.

1 | 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.

SDMs make several assumptions. One example is that species are at equilibrium with their environment (Zurell et al., 2020); otherwise, there will be habitats that are suitable for a species but in which it has not been recorded (because it does not occupy them). In this case, the model might erroneously predict that those habitats are unsuitable. This is just one example of how violations of a model's assumptions can affect its predictive performance.

The predictive 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 might 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). Species traits include range size relative to the study extent (Santika, 2011) and niche breadth (Hernandez et al., 2006; Tassarolo et al., 2021), among 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, and then assess the accuracy of those models.

Assessing the accuracy of a SDM generally involves comparing its predictions to data. These data might be the same data that were 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, AUC, Kappa and TSS have been criticized on several grounds. They measure rates of false presences and false absences so are challenging to interpret where presence-only datasets are used for evaluation. Of course, this limitation does not apply where presence-absence data are available (e.g. Phillips et al., 2009; Valavi et al., 2022). AUC, Kappa and TSS also depend on the ratio of presences to (pseudo-)absences, that is, sample prevalence, in the dataset (Jiménez-valverde et al., 2013; Leroy et al., 2018; Lobo et al., 2008). Alternative metrics have been developed to circumvent this latter issue (e.g. Kaymak et al., 2012), but we find that they are rarely used for SDM evaluation.

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; Tassarolo 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). Another limitation is that, as it is typically used—that is 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 recognized 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 predictive performance. We fitted SDMs for 1216 species of invertebrate and bryophyte in the United Kingdom (UK) using a fairly typical presence/pseudo-absence modelling workflow. We evaluated the performances of a subset (518; 43%) of these models, both in terms of variance among replicate model fits and accuracy as assessed by taxon experts. (For the speciose bryophytes, a random subset of 100 species were assessed, leaving 698 species unassessed.) We used DAGs to conceive plausible models describing the effects of explanatory variables on SDM performance and used multilevel path analysis to quantify those effects.

2 | METHODOLOGY

2.1 | 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 km² or finer between 2000 and 2015 to match the SDM covariate data (Appendix S1) and removed records that were duplicated in terms of grid cell and species (standard practice for species distribution modelling).

2.2 | Species distribution models

In this section, we briefly outline the SDM workflow (Figure 1), but refer the reader to the ODMAP (Overview, Data, Model, Assessment and Prediction; Zurell et al., 2020) document in Appendix S1 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 (cross-validation folds) 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 km² 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 15 replicate model fits.

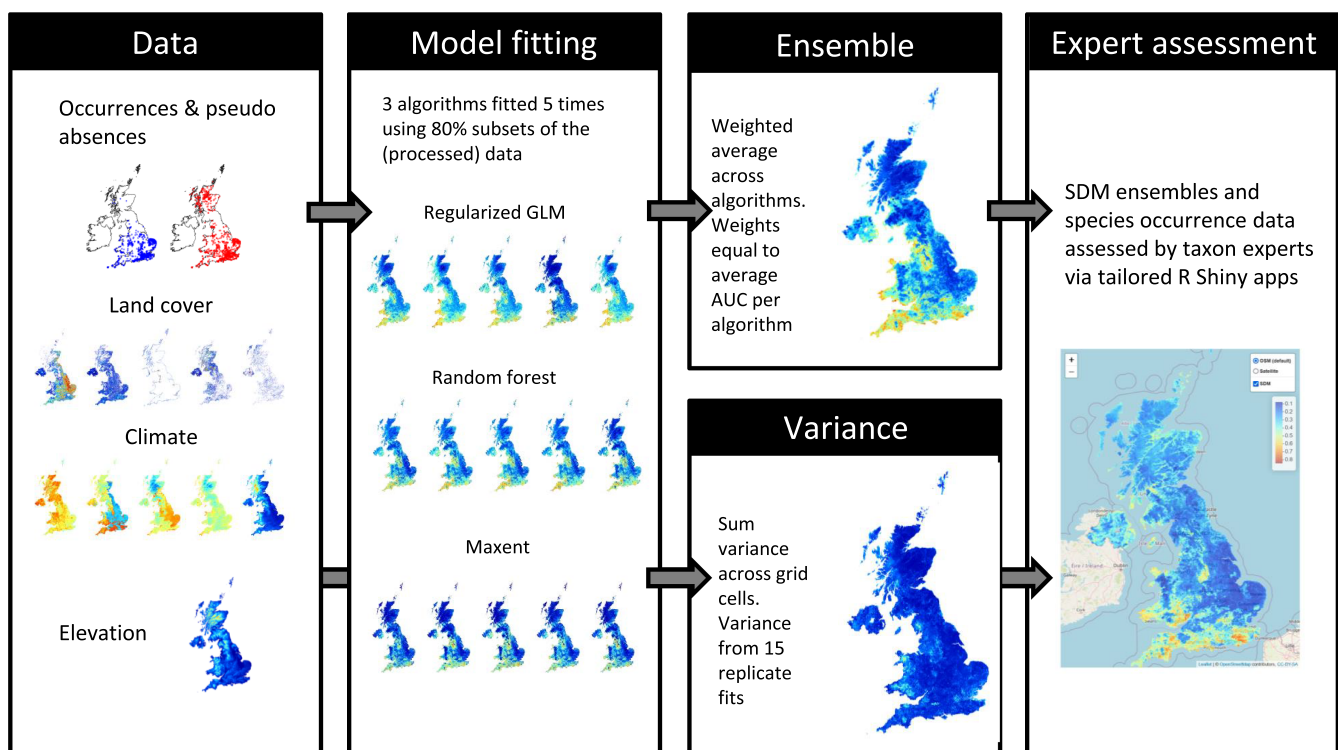


FIGURE 1 Species distribution modelling and assessment workflow. See the supplementary ODMAP document for full details (Appendix S1). Note that the SDM predictions were presented to the experts on a continuous scale (relative habitat suitability) rather than a binary one. Also note that the data subsetting in the "Model fitting" box was performed after the data were degraded to unique species-location combinations.

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* (Friedman et al., 2010), *randomForest* (Breiman et al., 2018) and *dismo* (Hijmans et al., 2017).

2.3 | 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). Among 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). We used five-point Likert scales to avoid creating a false sense of certainty about our results, having felt that it was unrealistic to expect that the experts could provide answers on a continuous scale (e.g. could one model be classed as 62% accurate and another 72%?).

Each expert was provided with a tailored R Shiny app, which included the predicted maps of habitat suitability (as a continuous measure), 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).

2.4 | 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). It 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 cross-validation folds, which are two major sources of variability in SDM predictions.

SDM algorithm contributes more than cross-validation fold to our measure of variance. To test this, we calculated mean habitat suitability across grid cells for each replicate fit. The median (across species) proportion of the total variance in mean habitat suitability explained by SDM algorithm is about 77%, compared to about 23% for cross-validation fold.

2.5 | Explanatory variables

We assumed that SDM accuracy and variance are functions of eight variables. Definitions of the variables, and information on if and how they were measured, are provided in Table 2.

TABLE 1 A taxonomic breakdown of the number of species modelled, the number of models assessed, the number of models “fully” assessed, the assessor initials (see author list) and their affiliations.

Taxonomic group	Number of species modelled	Number of species assessed	Number of species fully assessed	Expert initials	Recording scheme
Mosses, liverworts and hornworts (Bryophyta, Marchantiophyta, and Anthocerotophyta)	782	100	99	CDP	British Bryological Society (https://www.britishbryologicalsociety.org.uk/)
Centipedes (Chilopoda)	29	29	25	TB	British Myriapod and Isopod Group, Centipede Recording Scheme (https://www.bmig.org.uk/)
Dragonflies (Odonata)	46	46	45	PT	British Dragonfly Society Recording Scheme (https://british-dragonflies.org.uk/)
Hoverflies (Syrphidae)	226	226	208	RM	Dipterists Forum, Hoverfly Recording Scheme (http://hoverfly.uk/hrs/)
Mayflies (Ephemeroptera)	38	38	38	CM	Riverfly Recording Schemes: Ephemeroptera (http://www.ephemeroptera.org.uk/)
Soldierflies and allies (Lower Brachycera)	95	95	94	MH	Soldierflies and Allies Recording Scheme (http://soldierflies.brc.ac.uk/)
Total	1216	554	518	-	-

Note: By fully assessed, we mean SDMs for which the assessors could answer that all relevant questions. Data from the remainder of SDMs were not used to construct the causal model.

TABLE 2 Explanatory variables that we assume affect SDM accuracy and variance.

Variable	Definition	Derivation
Equilibrium	Degree to which species are at equilibrium with their environment.	Unobserved.
Niche breadth	Breadth of species' environmental niches (ranging from habitat specialist to habitat generalist).	Unobserved, but see Appendix S4 where we use a proxy measure.
Niche completeness	Degree to which the occurrence data cover a species' environmental niche.	Assessed by the experts on a five-point Likert scale.
Prevalence	Species' true range sizes.	Calculated as sample size divided by range completeness (1 being poor coverage and 5 being full coverage). Higher where sample size is high despite poor coverage of the species' range, and vice versa.
Range completeness	Degree to which the occurrence data cover a species' geographic range.	Assessed by the experts on a five-point Likert scale.
Recorder behaviour	Recorders' decisions about where to sample geographically (and hence environmentally).	Unobserved.
Sample size	The number of 1 km grid cells (EPSG:27700) in which the species has been recorded (between 2000 and 2015).	Empirical.

2.6 | 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 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.

Having started with one DAG, which was not empirically plausible, we tested a further nine (Appendix S2). The final DAG, which we present and analyse in this paper, is both theoretically plausible and empirically supported by our model/data combination. Full details of the model conceptualisation and testing process can be found in the R Markdown document in Appendix S2. Theoretical justifications for the effects posited by the causal model are laid out in Table 3.

2.7 | 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 standardized 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 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 $\rightarrow n \rightarrow$ accuracy in Figure 2. The total indirect effect of one variable on another is the sum of the specific indirect effects over all pathways linking them (Preacher & Hayes, 2008; Tarling, 2009).

A reviewer rightly pointed out that the effects listed in Table 3 might be nonlinear. Indeed, it was apparent on visual inspection of our data that sample size is nonlinearly related to SDM accuracy and variance. This is not surprising: others have noted an asymptotic relationship between sample size and SDM accuracy (Hallman & Robinson, 2020), and it is well known that the variance of an estimator (here the SDM ensemble) is an asymptotic function of sample size. To enable the use of path analysis, which is based on linear regression, we (natural) log transformed sample size before fitting the path models. It was also necessary to log transform prevalence to obtain a linear relationship between it and the log of sample size. Accuracy and variance are plotted against the log of sample size in Figure 3.

To assess the uncertainty associated with the effects estimated by the path model, we used nonparametric bootstrapping. We resampled the data by species (both response and explanatory variables) 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

TABLE 3 Theoretical justifications for the effects posited by our causal model. Justifications are stated *ceteris paribus*.

Explanatory variable	Response variable	Theoretical justification
Equilibrium	Niche completeness	A species cannot be recorded in portions of its niche that it does not occupy.
Equilibrium	Prevalence	For a given niche breadth, a species closer to equilibrium with its environment will be more widespread.
Niche breadth	Niche completeness	It is harder to sample a given portion of a generalists' than a specialists' niche.
Niche breadth	Prevalence	A generalist has the capacity to be more widespread than a specialist.
Prevalence	Range completeness	It is harder to sample a given portion of a widespread species' range than a common one.
Prevalence	Sample size	It is possible to record a widespread species in more locations than a rare one.
Recorder behaviour	Niche completeness	For a given niche breadth and equilibrium, the geographic locations sampled determine the fraction of a species' niche that is sampled.
Recorder behaviour	Range completeness	For a given prevalence, the geographic locations sampled determine the fraction of a species' range that is sampled.
Niche breadth	Accuracy	Pseudo-absences are likely to be placed in habitats that are suitable for generalists. Hence, their discrimination ability should be poorer than models for specialists.
Niche breadth	Variance	There is less scope for variation in the types of habitats that specialists are recorded in. This should reduce sampling variability.
Niche completeness	Accuracy	SDMs estimate species' environmental niches so adequate coverage of those niches is important.
Niche completeness	Variance	Maxent and random forests are more likely to find spurious signals where niche completeness is low (i.e. overfitting), compared to the regularized GLMs. This means that they will make different predictions in non-sampled portions of environmental space (Werkowska et al., 2017), increasing variability among algorithms.
Sample size	Accuracy	Larger samples will yield a more accurate point estimate in the absence of systematic error.
Sample size	Variance	Small samples are more likely to be unusual (different from the population) by chance, which increases sampling variability (Lohr, 2022).

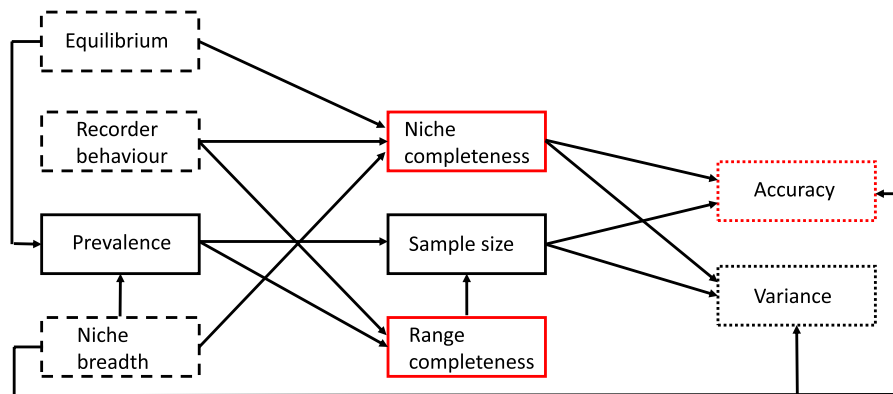


FIGURE 2 Directed acyclic graph (Greenland et al., 1999) depicting our causal model's assumptions about what determines SDM accuracy and variance. Dashed line boxes indicate unobserved variables, red boxes indicate expert-assessed variables and dotted line boxes denote the response variables of interest (Table 2).

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 of 0.08, 0.25 and 0.23 (Appendix S2, p. 41) indicate that the data are not

independent within assessors. Hence, we include a random intercept for assessor identity in the portions of our path model in which an expert-assessed variable is the response.

2.8 | 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 often 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 (Appendix S3), so we only present the results from the multilevel piecewise path models here.

3 | RESULTS

Our model explains 20% of the variation in SDM expert-assessed accuracy. Range completeness, prevalence, niche completeness and sample size have positive effects (Figure 4). The effects of range completeness and prevalence are indirect, whereas the effects of niche completeness and sample size are direct.

The model explains 59% of the variation in SDM model-based variance. Range completeness, prevalence, niche completeness and sample size have negative effects—as they increase, variance decreases (Figure 4). Like accuracy, the effects of range completeness and prevalence are indirect, whereas the effects of niche completeness and sample size are direct.

4 | 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. Cumulatively, they have considerable local, national and international field knowledge gained by writing distribution atlases, field guides, species status reviews and autecological papers. Some also undertake their own modelling using similar tools to those in this study. Hence, their assessments arguably reflect an unrivalled synthesis of information.

Although experts, there are limits to, and possibly biases in, the assessors' knowledge. In total, 554 species were assessed, but the experts were only able to provide answers to all of the relevant questions for 518. It is possible that these missing species (6.5%)

share common attributes that have biased our analysis (i.e. they are missing not at random; Rubin, 1976). While we have confidence in our experts' experience and knowledge for our area, it is of course possible that this does not generalize across all disciplines or datasets. In some cases, the available expert knowledge might still be biased towards particular places, such as areas of high population density or higher species richness, or particular latitudes. We do not claim that expert validation is a panacea for assessing SDMs: as always, the appropriateness of the tools and information available for making inferences need to be assessed for the case in hand.

Our model explains a relatively low proportion of the variation in SDM accuracy (20%; but see Møller & Jennions, 2002). To some extent, this is likely to reflect noise in the experts' assessments (of both accuracy and niche completeness). It might also reflect the fact that we missed important explanatory variables. One example is the appropriateness of the SDM covariates for a given species, which could affect model accuracy (Barry & Elith, 2006). We asked the assessors to report on this, but they felt unable to do so for the majority of species (56%). Another example is niche breadth, which we treated as unobserved. Omission of these variables could have biased the effects estimated by our model (Angrist & Pischke, 2009).

We suspect that the omission of niche breadth will not have appreciably biased our estimated effects. As a proxy for niche breadth, we calculated the number of land cover classes, using the UKCEH 2007 Land Cover Map (Morton et al., 2011), on which each species was recorded. This is not a perfect proxy for niche breadth, but we suspect that it will at least be a correlate thereof at the scale of our models (1 km²). Including this proxy measure of niche breadth in the models barely changed our estimated effects (Appendix S4). It is not clear, however, whether and to what extent our estimates would change if we were to identify, measure and include additional explanatory variables in our model.

Previous studies demonstrated positive effects of sample size and niche completeness on SDM accuracy (Feeley & Silman, 2011; Konowalik & Nosol, 2021; Stockwell & Peterson, 2002; Wisz et al., 2008), but not in a causal framework. A causal framework is needed, however, because there is a clear risk that sample size could confound the effect of niche completeness or vice versa. Indeed, two studies that reported correlations between sample size and SDM accuracy acknowledged that niche completeness was higher in larger samples (Feeley & Silman, 2011; Wisz et al., 2008). It would be interesting to know if the findings of these studies would have been less pronounced had niche completeness been conditioned on.

Our results suggest that niche completeness and sample size have similarly strong effects, which implies that the positive effect of sample size could be offset by low niche completeness. 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

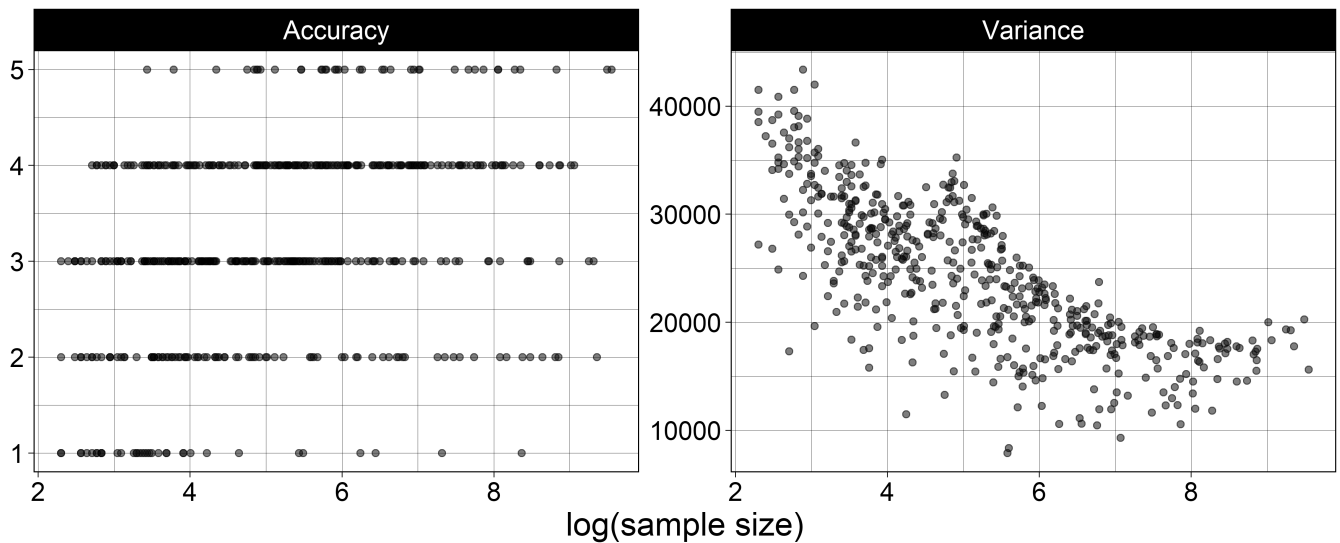


FIGURE 3 SDM accuracy and variance plotted against the log of sample size. Each point denotes one species. All variables are scaled and centred—after the log transformation in the case of sample size.

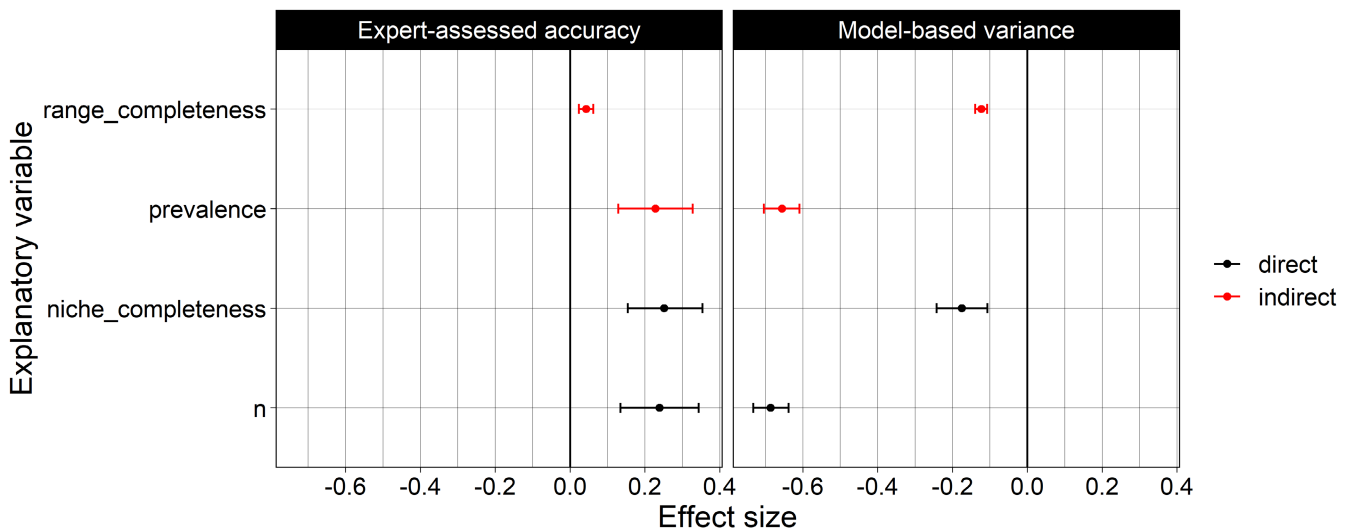


FIGURE 4 Effects of the explanatory variables on SDM expert-assessed accuracy and model-based variance. Effects are the z transformed path coefficients from Figure 2. The dots indicate the mean effect obtained by bootstrapping and error bars demarcate 95% (percentile) confidence intervals. The colour of the points and the error bars signify the type of effect: direct or (total) indirect.

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 to achieve this. One option is to consult appropriate 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 digitized range map—these

tools could be used to calculate niche coverage relative to the more complete data.

Another key insight from our analysis concerns the relationship between species' prevalence and SDM accuracy. Previous studies reported negative correlations between prevalence and accuracy conditional on sample size (Hernandez et al., 2006; Stockwell & Peterson, 2002; van Proosdij et al., 2016). It is important to remember, however, that larger sample sizes—when defined as the number of grid cells in which a species has been recorded—are possible for widespread species, which occupy more grid cells and are readily recognized by recorders with limited expertise. This is likely to be one reason why we found a positive indirect effect of species prevalence on accuracy (Figure 3).

Our model also suggests that prevalence has a negative indirect effect on SDM variance. Like accuracy, this effect is mediated by sample size. Hence, where sample size is correlated with prevalence, as in our dataset, more accurate and precise SDMs should be attainable for widespread species than rare ones. This notion challenges the conventional wisdom that rare species lend themselves better to modelling.

It should be noted, however, that our estimates of the indirect effects of prevalence on SDM accuracy and variance could be biased. Part of the total indirect effects of prevalence on accuracy and variance is the direct effect of prevalence on range completeness (Figure 2). Range completeness is also affected by recorder behaviour, which is unobserved. The omission of range completeness could bias the effect of prevalence on sample size, which could, in turn, bias the total indirect effects of prevalence on SDM accuracy and variance. Whether and to what extent this bias exists is unclear.

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.

We do not claim that our causal model is true. However, in depicting it as a DAG 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

We thank Diana Bowler and two anonymous reviewers whose comments improved this manuscript. We also thank the committee and members of the British Bryological Society, the British Myriapod and Isopod Group, the Centipede Recording Scheme, the British Dragonfly Society Recording Scheme, the Hoverfly Recording Scheme, the Riverfly Recording Schemes and the Soldierflies and Allies Recording Scheme. 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.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to disclose.

PEER REVIEW

The peer review history for this article is available at <https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/ddi.13698>.

DATA AVAILABILITY STATEMENT

The data and code underpinning this manuscript are available in the [Supporting Information](#).

ORCID

Robin J. Boyd  <https://orcid.org/0000-0002-7973-9865>

REFERENCES

- Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, 43, 1223–1232. <https://doi.org/10.1111/j.1365-2664.2006.01214.x>
- Amini Tehrani, N., Naimi, B., & Jaboyedoff, M. (2021). Modeling current and future species distribution of breeding birds as regional essential biodiversity variables (SD EBVs): A bird perspective in Swiss Alps. *Global Ecology and Conservation*, 27, e01596. <https://doi.org/10.1016/j.gecco.2021.e01596>
- Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University 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, 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, 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, 1173–1182. <https://doi.org/10.1037//0022-3514.51.6.1173>
- Barry, S., & Elith, J. (2006). Error and uncertainty in habitat models. *Journal of Applied Ecology*, 43, 413–423. <https://doi.org/10.1111/j.1365-2664.2006.01136.x>
- Bazzichetto, M., Lenoir, J., Da Re, D., Tordini, E., Rocchini, D., Malvasi, M., Vojtech, B., & Sperandii, M. G. (2022). Effect of sampling strategies on the response curves estimated by plant species distribution models. *EcoEvoRxiv*. <https://doi.org/10.32942/osf.io/rhys3>
- 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. J., Powney, G., Carvell, C., & Pescott, O. L. (2021). occAssess: An R package for assessing potential biases in species occurrence data. *Ecology and Evolution*, 11, 16177–16187. <https://doi.org/10.1002/ece3.8299>
- Breiman, L., Cutler, A., & Classification, D. (2018). Package ‘randomForest’. <https://cran.r-project.org/web/packages/randomForest/index.html>
- 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, 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, 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, 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, e0202403. <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*, 17, 323–330.
- Feeley, K. J., & Silman, M. R. (2011). Keep collecting: Accurate species distribution modelling requires more collections than previously thought. *Diversity and Distributions*, 17, 1132–1140. <https://doi.org/10.1111/j.1472-4642.2011.00813.x>
- Fourcade, Y., Engler, J., Rodder, 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, 1–13. <https://doi.org/10.1371/journal.pone.0097122>
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 128–129. <https://doi.org/10.1002/wics.10>
- 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>
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multi-level/hierarchical models*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511790942>
- Grace, J. B. (2006). *Structural equation modeling and natural systems*. Cambridge University Press.
- Grace, J. B., & Irvine, K. M. (2020). Scientist's guide to developing explanatory statistical models using causal analysis principles. *Ecology*, 101, 1–14. <https://doi.org/10.1002/ecy.2962>
- Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10, 37–48. <https://doi.org/10.1097/00001648-199901000-00008>
- Hallman, T. A., & Robinson, W. D. (2020). Deciphering ecology from statistical artefacts: Competing influence of sample size, prevalence and habitat specialization on species distribution models and how small evaluation datasets can inflate metrics of performance. *Diversity and Distributions*, 26, 315–328. <https://doi.org/10.1111/ddi.13030>
- Hao, T., Elith, J., Lahoz-Monfort, J. J., & Guillera-Aroita, G. (2020). Testing whether ensemble modelling is advantageous for maximising predictive performance of species distribution models. *Ecography*, 43, 549–558. <https://doi.org/10.1111/ecog.04890>
- Hernandez, P. A., Graham, C. H., Master, L. L., & Albert, D. L. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, 29, 773–785. <https://doi.org/10.1111/j.0906-7590.2006.04700.x>
- Hijmans, R. J., Phillips, S. J., Leathwick, J. R., & Elith, J. (2017). Dismo: Species distribution modeling. R Package version 1.1-4. <https://cran.r-project.org/web/packages/dismo/dismo.pdf>
- Hoveka, L. N., Bank, M., & Davies, T. J. (2022). Winners and losers in a changing climate: How will protected areas conserve red list species under climate change? *Diversity and Distributions*, 28, 782–792. <https://doi.org/10.1111/ddi.13488>
- Hoveka, L. N., van der Bank, M., & Davies, T. J. (2020). Evaluating the performance of a protected area network in South Africa and its implications for megadiverse countries. *Biological Conservation*, 248, 108577. <https://doi.org/10.1016/j.biocon.2020.108577>
- Jiménez-valverde, A., Acevedo, P., Barbosa, A. M., Lobo, J. M., & Real, R. (2013). Discrimination capacity in species distribution models depends on the representativeness of the environmental domain. *Global Ecology and Biogeography*, 508–516, 508–516. <https://doi.org/10.1111/geb.12007>
- Kaymak, U., Ben-David, A., & Potharst, R. (2012). The AUK: A simple alternative to the AUC. *Engineering Applications of Artificial Intelligence*, 25, 1082–1089. <https://doi.org/10.1016/j.engappai.2012.02.012>
- Konowalik, K., & Nosol, A. (2021). Evaluation metrics and validation of presence-only species distribution models based on distributional maps with varying coverage. *Scientific Reports*, 11, 1–15. <https://doi.org/10.1038/s41598-020-80062-1>
- Lefcheck, J. S. (2016). piecewiseSEM: Piecewise structural equation modelling in R for ecology, evolution, and systematics. *Methods in Ecology and Evolution*, 7, 573–579. <https://doi.org/10.1111/2041-210X.12512>
- Leroy, B., Delsol, R., Huguency, B., Meynard, C. N., Barhoumi, C., Barbet-Massin, M., & Bellard, C. (2018). Without quality presence-absence data, discrimination metrics such as TSS can be misleading measures of model performance. *Journal of Biogeography*, 45, 1994–2002. <https://doi.org/10.1111/jbi.13402>
- Lobo, J. M., Jiménez-valverde, A., & Real, R. (2008). AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17, 145–151. <https://doi.org/10.1111/j.1466-8238.2007.00358.x>
- Lohr, S. (2022). *Sampling: Design and analysis* (3rd ed.). CRC Press.
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7, 83–104. <https://doi.org/10.1037/1082-989X.7.1.83>
- McElreath, R. (2020). *Statistical rethinking: A bayesian course with examples in R and Stan*. Chapman & Hall/CRC. <https://doi.org/10.1201/9781315372495>
- Møller, A. P., & Jennions, M. D. (2002). How much variance can be explained by ecologists and evolutionary biologists? *Oecologia*, 132, 492–500. <https://doi.org/10.1007/s00442-002-0952-2>
- Morton, R. D., Rowland, C., Wood, C., Meek, L., Marston, G., Smith, G., Wadsworth, R., & Simpson, I. (2011). Land cover map 2007 (1km percentage target class, N. Ireland). <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, 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). <https://doi.org/10.5285/Oec7e549-57d4-4e2d-b2d3-2199e1578d84>
- 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*, 141, 1–7. <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, 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, 879–891. <https://doi.org/10.3758/BRM.40.3.879>
- R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581–592. <https://doi.org/10.1093/biomet/63.3.581>
- Santika, T. (2011). Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. *Global Ecology and Biogeography*, 20, 181–192. <https://doi.org/10.1111/j.1466-8238.2010.00581.x>
- Santini, L., Benítez, A., Huijbregts, M. A. J., Maiorano, L., & Čengić, M. (2021). Assessing the reliability of species distribution projections in climate change research. *Diversity and Distributions*, 40, 1035–1050. <https://doi.org/10.1111/ddi.13252>

- 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, 12858–12868. <https://doi.org/10.1002/ece3.5766>
- Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, 13, 290–312.
- Spiers, J. A., Oatham, M. P., Rostant, L. V., & Farrell, A. D. (2018). Applying species distribution modelling to improving conservation based decisions: A gap analysis of Trinidad and Tobago's endemic vascular plants. *Biodiversity and Conservation*, 27, 2931–2949. <https://doi.org/10.1007/s10531-018-1578-y>
- 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, 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*, 12, 216–226. <https://doi.org/10.1111/2041-210X.13525>
- Stockwell, D. R. B., & Peterson, A. T. (2002). Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, 148, 1–13. [https://doi.org/10.1016/S0304-3800\(01\)00388-X](https://doi.org/10.1016/S0304-3800(01)00388-X)
- Tarling, R. (2009). *Statistical modelling for social researchers: Principles and practices*. 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>
- Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J., & Elith, J. (2022). Predictive performance of presence-only species distribution models: A benchmark study with reproducible code. *Ecological Monographs*, 92, 1–27. <https://doi.org/10.1002/ecm.1486>
- van Proosdij, A. S. J., Sosef, M. S. M., Wieringa, J. J., & Raes, N. (2016). Minimum required number of specimen records to develop accurate species distribution models. *Ecography*, 39, 542–552. <https://doi.org/10.1111/ecog.01509>
- 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, 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, 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, 763–773. <https://doi.org/10.1111/j.1472-4642.2008.00482.x>
- Zellmer, A. J., Claisse, J. T., Williams, C. M., Schwab, S., & Pondella, D. J. (2019). Predicting optimal sites for ecosystem restoration using stacked-species distribution modeling. *Frontiers in Marine Science*, 6, 1–12. <https://doi.org/10.3389/fmars.2019.00003>
- 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, 1261–1277. <https://doi.org/10.1111/ecog.04960>

BIOSKETCH

Robin J. Boyd is a methodologist and ecologist. His main interest is in developing methods to draw inferences about plant and animal populations from big, unrepresentative data. You can find all Rob's research on his ResearchGate profile <https://www.researchgate.net/profile/Rob-Boyd-2>.

Author contributions: R.J.B.: Conceptualization (equal), methodology (lead), formal analysis (lead), riting—original draft (lead), writing—review & editing (lead) and visualization (lead). M.H.: Investigation (equal), writing—review & editing (supporting) and data curation (equal). D.B.R.: Project administration (lead) and writing—review & editing (supporting). T.B.: Investigation (equal). K.A.H.: Investigation (equal) and writing—review & editing (supporting). C.R.M.: Investigation (equal) and writing—review & editing (supporting). R.K.A.M.: Investigation (equal) and writing—review & editing (supporting). C.P.: Investigation (equal). S.P.: Investigation (equal). C.D.P.: Investigation (equal) and writing—review & editing (supporting). P.T.: Investigation (equal). R.W.: Investigation (equal) and writing—review & editing (supporting). S.G.B.: Investigation (equal). O.L.P.: Conceptualization (equal), methodology (supporting), formal analysis (supporting), writing—original draft (supporting), writing—review & editing (supporting) and data curation (equal).

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Boyd, R. J., Harvey, M., Roy, D. B., Barber, T., Haysom, K. A., Macadam, C. R., Morris, R. K. A., Palmer, C., Palmer, S., Preston, C. D., Taylor, P., Ward, R., Ball, S. G., & Pescott, O. L. (2023). Causal inference and large-scale expert validation shed light on the drivers of SDM accuracy and variance. *Diversity and Distributions*, 29, 774–784. <https://doi.org/10.1111/ddi.13698>