Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/5/1/eaat4858/DC1

Supplementary Materials for

Standards for distribution models in biodiversity assessments

Miguel B. Araújo*, Robert P. Anderson, A. Márcia Barbosa, Colin M. Beale, Carsten F. Dormann, Regan Early,
Raquel A. Garcia, Antoine Guisan, Luigi Maiorano, Babak Naimi, Robert B. O'Hara,
Niklaus E. Zimmermann, Carsten Rahbek

*Corresponding author. Email: maraujo@mncn.csic.es

Published 16 January 2019, *Sci. Adv.* **5**, eaat4858 (2019)
DOI: 10.1126/sciadv.aat4858

This PDF file includes:

Supplementary Text

Text S1. Uses of models in biodiversity assessments.

We reviewed how correlative Species Distribution Models (SDMs) have been applied in biodiversity assessment by searching the Thomson Reuters' Web of Science core collection on 18.05.2016 for peer-reviewed articles using SDMs in a biodiversity context. As search terms, we used a range of alternative names to describe correlative SDMs, combined with terms relating to biodiversity to filter out non-biodiversity applications (table S1.1). We restricted the results to articles published from 1995 to 2015, written in English, and in journals from the ISI ecology and biodiversity conservation lists. Our search yielded 6,483 articles.

We randomly sampled 400 papers to review (~6% of all papers retrieved). The vast majority of the papers in our sample modelled between one and five species (fig. S1.1), with terrestrial habitats in North America or Europe receiving most attention (fig. S1.2). We classified the papers according to the purpose for which the SDM was built (explanation, prediction or projection; table S1.2), and the intended conservation application of the SDM (table S1.3). Papers that developed or tested SDM methods were also marked as 'Methods', and we also noted any review papers. Papers that did not construct and apply SDMs were excluded from the classifications. To ensure that our sample was representative, we built accumulation curves showing the accumulated percentages of papers falling in each of the classes in table S1.2 and S1.3, as well as in different taxonomic groups, continents (fig. S1.4), and ecological realms, at increasing sample sizes, in increments of forty papers (fig. S1.3).

Text S2. Guidelines for scoring models in biodiversity assessments.

We developed four tables describing the guidelines to score SDMs for biodiversity assessment according to the standards set out in the main text. The tables relate to: 1) data used as a response variable (typically species occurrence data); 2) predictor variables (typically environmental data); 3) model building; and 4) model evaluation. For each of the four issues covered, we provide descriptions of the criteria used to score levels of quality according to our standards: aspirational (Gold); current best practice

(Silver); current minimum acceptable practice (Bronze); and current unacceptable practice (Deficient). We supply these descriptions for three different possible purposes (explanation, prediction, projection) for a given study (table S1.2).

We note two important caveats regarding these guidelines, especially for the purpose of Explanation and Projection. These caveats relate to the geographic extent of the study, and to differences between experimental vs. correlative approaches. First, we constructed these guidelines for studies that aim to characterize the drivers of distributions across the entire geographic or environmental range of the species. The guidelines may also prove relevant for studies within a smaller portion of a species' geographic or environmental range. In these cases it is possible that some standards for 'explanation' and 'projection' may be disregarded without compromising the quality of the SDMs. We note this in the table. Second, we fully acknowledge that firm conclusions regarding processes cannot be reached solely via correlative analyses of observational data. Rather, we note the ultimate need to consider jointly the results of both experimental and correlative approaches. We suggest caution in causally interpreting results from correlative analyses.

Below, we provide a short explanatory text to accompany the tables S2.1–4. The guidelines in each cell of a table provide guiding principles rather than detailed methods. The explanatory text adds brief examples of specific implementations. In the accompanying text, we include citations from the scientific literature, both to document the specific recommendations and to point the user toward helpful references. We anticipate that the implementation methods will change rapidly over time. We also note that when final results arise from an ensembles modeling approach using multiple SDM techniques (see glossary and 3D for definition), then the standards can only be evaluated by assessing both the constituent SDMs and ensemble.

Text S2.1. Guidelines for the response variable.

Summary: The response variable is typically the place (and time) that a taxon is known to occur or not to occur, although abundance or other population-level information (e.g., growth rate) are also sometimes used. Such data should be representative of both the spatial and the environmental distributions of the taxon, so that models can capture relevant relationships with the environment and predict and project trends regarding the species spatial distributions. The description of the guidelines for evaluating the biological response data used in SDMs is synthesised in table S2.1 and discussed below.

Text S2.1A. Sampling of response variables.

Repeated systematic surveys with sufficient and well-designed sampling that is quantified and evenly distributed (geographically and/or environmentally) are gold standard. Non-systematic surveys (e.g., affected by subjective choices) and imperfect species detection can have significant negative impacts on model accuracy, especially if resulting biases are spatially and/or environmentally structured (*85*). In such cases, incorporating information on species detectability and survey effort can improve modelling results (*86, 87*) (see also 1.2.3, 3B)· Lacking such information, other processing must be undertaken to reduce bias and better approximate geographically and environmentally representative samples (*56, 88*) (see also 1D and 1E)·

Text S2.1B. Identification of taxa.

Realistic models require taxonomic identifications that are rigorous and well documented. False presences due to misidentification can affect models adversely (*89*). The gold standard requires that identifications be provided by experts using multiple lines of evidence and based on records that can be re-examined (e.g., voucher specimens, DNA sequences, sound recordings, photographs). When this is not feasible and/or data come from heterogeneous sources, uncertainty regarding identification should be acknowledged and dealt with (e.g., by re-examining dubious records and/or by identifying and removing implausible ones) (*90*) (see also 1C).

Text S2.1C. Spatial accuracy of response variable.

For the gold standard the spatial accuracy of the response variable should match or be finer than the spatial resolution of the predictor variables employed. To allow appropriate treatment of the response variable, data should be geo-referenced soundly with both precision and accuracy being reported (ideally with a GPS) (*37, 91*). If a species is surveyed on a grid of cells of a certain size, the environmental variables should ideally be measured at, or (dis)aggregated to, cells of the same size, or methods should be used for dealing with data that are spatially discordant (see 2B). When the precision and accuracy of occurrence data are not known or are inconsistently quantified, implausible locations (e.g., species records outside a reasonable occurrence area) should be excluded from the model. Deficient practices include ignoring the spatial accuracy of species occurrences and using all available data blindly.

Text S2.1D. Environmental extent across which response variable is sampled.

For models to capture the species-environment relationships correctly and completely, data should include records across the entire range of environmental conditions suitable for the taxon (gold standard). Otherwise, models will be environmentally truncated (*40, 92*). When possible, multiple lines of evidence will be used to ensure that the species full environmental tolerances are modelled. Evidence could include environmental associations (and related environmental extents) drawn from species' past distributions or locations where species have naturalised outside their native range. This is especially important if the model is to be used for projecting outside the modelled area/time or for explaining the factors driving the distribution (*93, 94*). Data regarding past associations may be inferred from palaeoclimatic reconstructions or simulations, together with evidence of historical or pre-historical occurrence (e.g., fossil records or genetic evidence of long-term persistence) (*41, 95-98*). The data should ideally include the full variety of available environments and biotic contexts in which species occurs (*99, 100*). When there is no evidence that the taxon's full environmental tolerances are included, extrapolation of the model to conditions outside the environmental ranges of the variables used in model building must be avoided or acknowledged, and its potential effects discussed (*92*).

Text S2.1E. Geographic extent across which response variable is sampled (includes occurrence data and absence, pseudo-absence, or background data).

The extent of the study area can strongly affect modelling results (see also 1C)' so the region for absence/pseudo-absence/background samples should include but not exceed all areas that are accessible to the taxon (*101*). If the extent is too small, the importance of broad-scale range determinants such as climate may be underestimated; if the scale is too large, the model may not be able to capture important local nuances and instead reflect spurious environmental correlations with dispersal barriers (*98, 101*). The gold standard requires that the geographic extent be defined specifically for the focal taxon (e.g., based on dispersal ability and relevant environmental barriers), and including both the taxon's current and historical distribution. When this is not possible, reasonable extents can be approximated based on the current range using species-specific information and/or predefined biogeographic regions.

Text S2.2. Guidelines for the predictor variables.

Summary: Predictor variables are used to assess species-environment relationships (or to impose an a priori known relationship in mechanistic models), which, in turn, are often projected back to geographic space in order to map potential species distributions and biodiversity patterns. The acquisition, preparation and selection of predictor variables are thus crucial steps for construction of species distributions models. The gold standard requires that the conditions on which the response variable depends be captured by the predictor variables selected, at the spatial and temporal resolutions chosen, and that the uncertainty around these conditions (i.e., based on measurement and/or model errors) be quantifiable in the final model. The description of the guidelines for evaluating the predictor (often environmental) data used in biodiversity models is synthesised in table S2.2 and discussed below.

Text S2.2A. Selection of candidate variables.

A large number of candidate variables are usually available, so those selected for input into SDM analyses should ideally be causally related to a species' distribution. The gold standard is to use proximal variables exclusively (*36, 48, 102*), for which the effect on a species' distribution is well evidenced, so that a model builds on true cause-and-effect relationships. Variables should include all relevant proximal predictors (i.e., not being restricted to readily available climatic variables). Distal variables that are only associated with the species' distribution by correlation (i.e., indirect variables such as elevation or latitude) should not be used for models seeking explanation and projection (*36, 103*). In the absence of knowledge on proximal predictors, the selection of variables could be statistically justified using, for example, path analysis (*104*) or variance partitioning (*105, 106*), testing alternative predictor variable subsets if the exact causal set is not known with certainty (*50, 51*), and excluding spurious relationships that can appear under repeated testing (*107, 108*). Biotic variables (e.g., distributions of competitors or prey) should also be included whenever possible (*109-111*) and when there is evidence that interactions are strong (*23*) and impactful at the scale that is relevant for the inferences being made (*112*).

Text S2.2B. Spatial and temporal resolution of predictor variables.

The gold standard for the spatial and temporal resolution of predictor variables is to match the spatial resolution at which the response variable is affected by the predictors [see 1D]. Choosing an inappropriate spatial resolution can obscure the species–environment relationship and result in reduced predictive ability (*113, 114*). In addition, temporal resolution should match the biology of the response variable. For example, for seasonally migrating species, predictor variables should provide information throughout the year, rather than annual averages (*115*). Similarly, because the spatial distributions of many taxa show delayed responses to changes in predictor variables (*116-120*), analysis of historical time periods may provide better insights (*94, 96, 97*). Adjusting the spatial resolution of predictor variables to an appropriate (theoretically justified) resolution by downscaling or interpolation (*64*) is only acceptable when direct measurements are unavailable.

Text S2.2C. Uncertainty in predictor variables.

Uncertainty regarding the measurement and/or choice of the predictor variables entering the models can obscure the species–environment relationship and result in reduced predictive ability (*50, 51*). Gold-standard practice is to quantify the effects of the relevant uncertainties of predictor variables on SDM outputs (see 3D) (*57, 121-123*). Quantifying the effect of uncertainty from the choice of SDMs and climate models and/or greenhouse-gas emission scenarios is common when projecting species future ranges by means of ensemble modelling (*29, 58, 59, 124*) and more rarely by applying error propagation techniques to generate predictions that reflect predictor variable or other uncertainties (see 3D). Similar analysis of uncertainty should be done for other variable types (e.g., land use, soil conditions, biotic interactions). When uncertainty in predictor variables is not quantified, studies should minimally acknowledge likely sources (*125*) and consider their possible effects in interpretations (*126*).

Text S2.3. Guidelines for model building.

Summary: Model building includes fitting a correlative relationship between biological occurrence data and predictor variables. It includes consideration of model complexity and procedures to take into account imperfections in response and predictor variables. These considerations commonly include the effects of unrepresentative biological response data (e.g., due to biased sampling) and, more generally, characterization of uncertainty in model outputs resulting from various sources (both in the data and inherent to the model building process itself). The description of the guidelines for evaluating model building used in biodiversity models is synthetized in table S2.3 and discussed below.

Text S2.3A. Model complexity.

Complexity can include the number of parameters used in the models, flexibility of the modelling approach to fit them, and the number of operations involved (*54*). Formally, it can be quantified with metrics of computational complexity (*55*), defined by the amount of computational resources required to produce an output (*127*). Although more complex models will tend to explain training data better (*53, 128*), they may over-fit and do a poorer job at predicting and projecting (*129-131*). It has therefore long been recognised that any assessment of the effects of model complexity will depend on the purpose of the model (*132*), but the divide between prediction and projection may be equally important for biodiversity assessments. The optimal level of complexity required can be difficult to assess, especially when projecting to new conditions (in time or space) (*55, 133*). The gold standard is that model complexity and over-fitting are assessed in multiple ways, with methods such as penalization or model selection being employed when the aim is prediction. Complexity is best assessed by comparing inferences to fully independent data sets (see extended discussion in the next section 1.2.4, 4B) but, as a lesser standard, internal cross-validation can be used (*134*).

Text S2.3B. Treatment of bias and noise in response variables.

Biological response data may be biased or not sufficiently complete (See 1A). The gold standard requires either evidence for lack of bias, or that bias is addressed fully by methods including an observation model (e.g., an occupancy model) (*135*), or by including covariates that explicitly control for the bias (*136*). As a gold standard, the efficacy of the bias correction can be assessed by comparison with independent data, such as from high quality surveys in different regions or times. The silver standard approach to bias correction weights the data by their reliability, and assesses the result through internal cross-validation. Minimally, likely biases are acknowledged and described, along with a discussion of the effects that they may have on interpretations.

Text S2.3C. Treatment of collinearity.

Collinearity can cause problems in model fitting: if two variables are strongly co-linear, or correlated, it can be impossible to separate their effects. This is problematic when predicting to sites or environments in which these correlations change. The estimates of the effects of correlated variables can be very unstable, and projection uncertainty can increase hugely (*103*). Reviews (*103*) document the effects of correlation among predictors on the outcome of biodiversity models. Removing correlated variables is one solution, but the choice of which variables to remove has to be well informed ecologically, ideally via external information regarding likely causal mechanisms (see table S2.2, Predictor variables). If collinear variables are retained, the model building process (see 3A, model complexity) should aim at stabilising estimates (*103*). The gold standard would be no collinearity among variables. When collinearity cannot be removed, a model structure informed by ecology, rather than statistical information criteria, would be next-to-ideal. The silver standard is to use methods that are insensitive to collinearity or to stabilise the model in the presence of collinearity.

Text S2.3D. Dealing with modeling and parameter uncertainty.

It is important to estimate both the amount of uncertainty in the models and the parameter values, and to evaluate how these sources of uncertainty affect model outputs (e.g., whether uncertainty is particularly high in some parts of the species' range). The gold standard is that all relevant sources of uncertainty are assessed and incorporated into the analysis. One approach to dealing with uncertainty in parameterising species-environment relationships is to calculate standard errors and confidence intervals. Uncertainty arising from alternative model structures (e.g., multiple SDM approaches) can be averaged, in which case the weights given to the models need to be carefully considered (*57*). The gold standard is to fully address the sensitivity to uncertainties in models and parameters (e.g., data, model choice, initial conditions in any simulations, and parameters). Such process can be achieved by error propagation (*58*, *123*) in which uncertainty in all major parts of the models are reflected in the final predictions (*57*). Useful error propagation techniques are bootstrapping the entire analysis (which takes into account model uncertainty), Monte-Carlo simulations from parameter distributions (in particular for stochastic process models), analytical error propagation (for deterministic models), and description of the full likelihood (as in Bayesian model estimation). A silver standard would entail addressing at least the most relevant sources of uncertainty, e.g., based on literature evidence.

Text S2.4. Guidelines for model evaluation.

Summary: Models are expected to approximate ecological reality, and should be evaluated against data that are representative of the spatial, temporal, and environmental distributions of the response variable. Ideally, such data should be statistically independent from that used for model building. The description of the guidelines for appraising model evaluation approaches in biodiversity models is synthesised in table S2.3 and discussed below.

Text S2.4A. Evaluation of model assumptions.

Violation of the theoretical and statistical assumptions of a particular model can lead to unreliable results (*36, 134, 137, 138*) for model interpretation, geographic predictions, and projections (*37, 49*). Demonstrating that no model assumptions were violated is a gold standard in modelling. In cases where a researcher tests assumptions and finds departures from them, it is necessary to assess the consequences on interpretation of the results. If violation of assumptions cannot be avoided, explicit exploration and discussion of consequences for the interpretation of results in the particular context in which they are being used represents the bronze standard (*24, 49*). Blindly using models without testing assumptions should be considered a deficient practice.

Text S2.4B. Evaluation of model outputs.

A key step in evaluating models is to compare their results against statistically independent data (*134*). This generally means using a testing dataset of response and predictor variables that are spatially or temporally independent from the training dataset to avoid artefactual inflation of performance measures (*129, 139, 140*)· The best currently available independent data come from separate geographic areas (*130, 140-142*), or from data collected in different time periods than the ones used to construct the models; cases of temporally independent test data for evaluation of model outputs have included repeated species inventories over time (*27*), fossil data to test the ability of models to project distributions in different times (*96, 97*), or ancient genetic data to test for the ability of models to project past climatic suitability as a surrogate for abundance (*143, 144*). An alternative line of independent evidence is to experimentally evaluate model results, for example, using transplant experiments (*145*). The gold standard is to evaluate model results using multiple lines of independent evidence. If fully independent data are unavailable, evaluating model projections, or transferability, across space can be done by spatially, rather than randomly, sub-setting the data (*139, 146-148*). When sub-setting is not possible (e.g., due to small sample sizes), evaluation by repeated sub-sampling the training data in spatial blocks may be acceptable, although the approach should be used to verify the ability of the model to estimate the training data rather than to enable statements about the models' ability to project to new time periods or locations.

Text S2.4C. Measures of model performance.

The performance of a model can be assessed from many different perspectives (*149-154*). The traditional statistical approach to predictive performance is to quantify how close the predictions are to the actual outcome, using goodness-of-fit statistics (e.g., R2 statistics) (*155*) and other measures of agreement between predictions and independent observations (*149, 156*). For probability predictions from presence-absence or presence-only models, performance assessment includes estimating accuracy, bias, calibration, discrimination, refinement, resolution, and skill (*157*), as well as characterising spatial, temporal, and environmental patterns in errors (*126*). Reporting several performance measures that reflect different aspects of a model's performance (gold and silver) (*158*) is superior to reporting only one aspect (bronze). When probabilistic predictions are compared to presence-absence or presence-only observations, often only two classes of measures are used: calibration and discrimination. Calibration here can be defined as "the extent to which a model correctly predicts conditional probability of presence", and includes parametric measures such as calibration plots (*153*) and the Boyce index (*159*). Discrimination is "the ability to distinguish between occupied and unoccupied sites" and includes the non-parametric statistics of sensitivity, specificity, and the area under a receiver operating characteristic curve (*149*). Furthermore, whenever possible, the spatial, temporal and environmental pattern of errors and variance should be comprehensively characterised (*160, 161*), as these can change the interpretation of model predictions and related conservation decisions (*162*).

Text S3. Scoring a representative sample of the literature according to the guidelines.

To select papers for scoring, we started with the literature search described in supplementary text S1, but then restricted the journal list to those having a 5-year impact factor above 2.5 in the ISI ecology and biodiversity conservation lists (table S3.1). The search yielded 5140 papers. We randomly selected a pool of 700 papers from the initial list, and then removed all papers that were not appropriate for our analysis (e.g., papers that were purely methodological and did not apply models to a biodiversity scenario, or papers that mentioned but did not actually construct models). We obtained a final set of 400 papers that were scored by six authors (BN, LM, CFD, AMB, CB). The results of the scoring of the 400 studies are plotted in fig. S1.5 and discussed in the main text.

To provide a measure of the accuracy in the scoring of the papers, 80 of these were randomly selected and independently re-evaluated by one of the three different authors (AG, BOH, NZ). In the first step, for each paper, we measured the difference between the scores recorded in the two evaluations for each criterion. We then took the absolute of these measures, as the absolute errors, that ranged between a minimum of 0 (i.e., no difference), and a maximum of 3 (i.e., maximum possible differences) (fig. S1.6). The errors were then standardised by dividing to 3, thus they ranged between 0 and 1 (fig. S.1.7). We then quantified the mean errors for each criterion for each year.

Then to test if, despite variation across the quality of the different modelling studies, there were consistent differences of quality over time, we fitted a cumulative logistic mixed model to the data. This is an extension of a proportional odds logistic regression. So, for classification j in study i, the log odds of being above class k or better is

$$\log(q_{ijk}/(1-q_{ijk})) = \alpha_{jk} - \beta_{jt_i} + \varepsilon_{it}$$

log(q_ijk/(1-q_ijk)) = \alpha_jk - \beta_j t_i + \epsilon_it

where α_{jk} is the intercept for being in class k or lower for classification j, β_j is the trend in classification over time for year t_i of study i (and is of primary interest: a positive value means that a higher class is more likely), and ε_{it} is a random effect for class i at time t (note that we have replicate studies in each year). We assume that \epsilon_it ~ N(0, \sigma_\epsilon^2) $\varepsilon_{it} \sim N(0, \sum \varepsilon^2)$, and \beta_j ~ N(0, \sigma_\beta^2) $\beta_j \sim N(0, \sum \beta^2)$.

Because the model is the probability of being in class k or lower, there is a constraint that \alpha_jk < \alpha_jk+1 $\alpha_{jk} < \alpha_{jk+1}$. Also note that there are k-1 \alpha_i's α_i's.

We fitted the model with a Bayesian approach. We assumed uniform prior distributions between -1000 and 1000 for \alpha_jk α_{jk}, subject to the ordering constraint. We assume uniform distributions between 0 and 1000 for the hyper-parameters, i.e., the standard deviations \sigma_\epsilon $\sum \varepsilon$ and \sigma_\beta $\sum \beta$. The model was fitted with OpenBUGS3.2.2, through the BRugs package (*163*). Three chains were run and after a burn-in of 1000 iterations a further 5000 iterations were sampled.

The estimates obtained can be interpreted approximately as the change in probability that a modelling study will be in a higher class. So, for example, there is about a 3.4% higher probability that a study will be one class higher in model building if it was published in 2015 rather than 2014 (Figure 4). More generally, the analysis shows there is evidence of improvements in model building and model evaluation over time and, to some extent, in the response variables, but not so in the handling of predictor variables.

Text S4. Glossary

Glossary

Absence data - Datasets containing "records" of places where sampling has occurred but the taxon has not been documented. Typically used to characterize information on the predictor variables found at such sites (and compare this with information for sites where the taxon's presence has been documented).

Accessible area - The geographic regions that have been accessible to the taxon within the time span for which the response-environment variable relationship is to be measured, and for which the response and predictor data were collected. For example, the area to which a taxon could have occurred within recent generations or since the

Last Glacial Maximum. An accessible area is one in which the taxon has not been prevented from occupying given the taxon's dispersal ability and the configuration of barriers not included as a predictor variable, such as mountain ranges.

Accuracy - The degree to which data or model prediction/projection match the reality of a taxon's.

Algorithm - A set of mathematical rules that can be followed to produce an outcome. For example, iterative weighted least squares is an algorithm that can be used to find the maximum likelihood solution for a GLM. In the literature, the term has sometimes been incorrectly confused with the models that are to be fitted.

Algorithmic settings - The settings of an algorithm (here, for modelling the response–predictor relationship). The settings (either default, or deviations from them) will determine how efficiently the eventual parameters of the model are estimated and, in some cases, even the values of those parameters. Note, see difference between algorithmic settings and model parameterisations.

Background data - Data sets for places across the study area (the 'background'), whether or not sampling has occurred, and whether or not the taxon of interest has been found. Typically used to characterize predictor variables of such sites (and compare it with that for sites where the taxon's presence has been documented). Contrast with Pseudo-absence data.

Bias (in response data)– The systematic variation in the probability that sites (or types of sites) have been sampled. Often, such bias corresponds to accessibility for researchers (in geographic space) and often it also leads to sampling bias in environmental space.

Bias (of a statistic or model) - A systematic (i.e., directional) difference between the true value of a statistic and the estimate. Contrast with Sampling bias.

Biotic variables - Data on taxa whose distributions affect the distribution of the focal taxon (see Predictor variables and Environmental tolerances).

Candidate variables - A set of predictor variables from which a subset is selected for input into the algorithm (which will yield a model by fitting parameter values for the predictor variables).

Data uncertainty - (see Uncertainty analysis).

Discrimination - A non-parametric characterization of the ability of a model to correctly order (i.e., rank) positive instances higher than negative ones (here suitable vs. unsuitable; or present vs. absent).

Distal variables - Proximal and distal refer to the position of the predictor in the chain of processes that link the predictor to its impact on the organism of the focal species. A distal variable is only linked to the proximal variable (either causally or even more distally via non-causal correlation).

Ensembles – ensembles of models are obtained by generating multiple simulations (copies) across more than one set of initial conditions, model classes, parameters and boundary conditions. The different simulations may or may not be combined to produce a single composite estimate. When a single composite estimate is obtained it is often termed consensus.

Environmental data - See Predictor variables.

Environmental tolerances - The environmental requirements (abiotic and biotic) that need to be fulfilled for a population to survive.

Error propagation - Allowing the effects in all parts of a model to flow through to the final result. For example, uncertainty in predictors creates more uncertainty (and bias) in the estimated effects of these predictors, and thus should affect the uncertainty in the ultimate predictions and projections.

Explanation - Exploring statistical relationships between response and predictor variables, as a means of generating and testing hypotheses regarding a species' relationships with the environment.

Independent data - Data sets that are statistically independent (e.g., test datasets showing no spatiotemporal correlations with the training data used to build the model). The concept here applies both to occurrence records and values for environmental predictors. Fully independent data seldom exist, but spatially and/or temporally distinct data partitions often provide a high degree of independence. Temporally distinct data partitions could include observations of range shifts in recent decades. In contrast, random splits of a dataset into training and testing data are correlated, and therefore non-independent.

Model - Here, a mathematical description of how the predictor variables affect the response variable. Thus, a GLM is one model and a neural networks another.

Model classes - Types of models as defined by the functions or rules used to fit or construct them. Some examples include distance-based envelope methods (BIOCLIM, DOMAIN), regression-based approaches (e.g., general linear or additive models), and classification trees.

Model complexity – Most often loosely defined. Term sometimes used to mean the effective number of degrees of freedom. It has been equated to dimensionality (the number of predictor variables used in a model) and the amount of computational resources required to producing a given output (also known as computational complexity.

Model uncertainty - (see Uncertainty analysis).

Model parameterisations - Statistical models typically have parameters (such as a and b in the linear regression model y = a + bx) that are estimated from the data. In contrast, variables (e.g., y and x) are the entities that the model aims to represent or predict. A model parameterisation is one particular set of parameters. Note, see difference between model parameterisations and algorithmic settings (which lead to a model with certain parameter values).

Multiple lines of evidence - Evidence from different types of data, supporting a similar posit or inference. Examples include fossil records, spatially distinct areas (such as native and naturalised ranges), spatial patterns of genetic diversity, and manipulative experiments.

Noise (in response data) - Random variation, without bias or consistent signal. Contrast with Bias.

Non-analogue – Values of predictor variables that lie beyond the range of those found in the dataset used for building model. This can be for one or more predictor variables, or for combinations of them. Such conditions typically occur when models are applied (projected/transferred) to news regions or time periods.

Occurrence data (a type of response variable) - Records of a taxon’s presence (and sometimes abundance), typically from specimens in natural history museums and herbaria and/or from visual and auditory observations.

Prediction - Quantified statement made by a parameterized model. Predictions to an environmental domain within the range of predictor values used for model parameterisation is called interpolation. When predictions are made beyond the environmental domain of the range of predictor values used to parameterise the model, it is called extrapolation. See also: projection

Predictor variables - Variables that are used in an algorithm to build a model that predicts the response variables.

Projection – Specific type of prediction in which environmental suitability for a taxon (its potential geographic distribution) is estimated in a different time period or region from the data used to construct the model, often involving extrapolation. Also called ‘transferring’.

Proximal variables - Proximal and distal refer to the position of the predictor in the chain of processes that link the predictor to its impact on the organism of the focal species. A proximal variable determines the organism’s response. Contrast with Distal variables.

Pseudo-absence data - Datasets containing ‘records’ of places where a taxon has not been observed, whether or not sampling occurred. Note that the taxon may actually inhabit such sites, but not be recorded as present, due to inadequate or non-existent sampling. Typically used to characterise environmental information of such sites (and compare it with that for sites where the taxon’s presence has been documented).

Positional error – Also known as geo-referencing error or location error. Error in spatial location of data (here, either occurrence data or environmental variables), for example due to recording uncertainty, misalignment of datasets, or changing of spatial grain.

Resolution (spatial) - The size of the cells, sometimes called pixels, of the raster grid in the study region in geographic space (= grain).

Response variable - Includes occurrence data, but also geographic records of a taxon’s abundance and population-level information (e.g., growth rate).

Sensitivity analysis - Quantification of the effect of changing model parameters on the model output. Aims at identifying most uncertain inputs or relationships.

Sub-sampling of training data - A partitioning (or splitting) of occurrence data (e.g., repeated random splits into training and testing datasets).

Time error - Error in temporal “location” of data (here, either response or predictor variables), for example due to recording uncertainty, misalignment of datasets, or changing of temporal grain.

Theoretically justified - Justified by ecological theory, rather than empirical data specific to the taxa involved.

Training data - The subset of the response variable used to build the model.

Uncertainty analysis - Quantification of the effect of uncertainty in any step of the analysis (data, model structure, parameterization, scenarios, ideally including their interaction through error propagation) on the model predictions.

Unreasonable occurrence records - Those occurrence records considered biologically or factually implausible by the researcher (e.g. records far from the documented range of the taxon, in the sea for a terrestrial species, or geo-reference that falls into a country different from that indicated in the ‘country’ field).

Supplementary figures

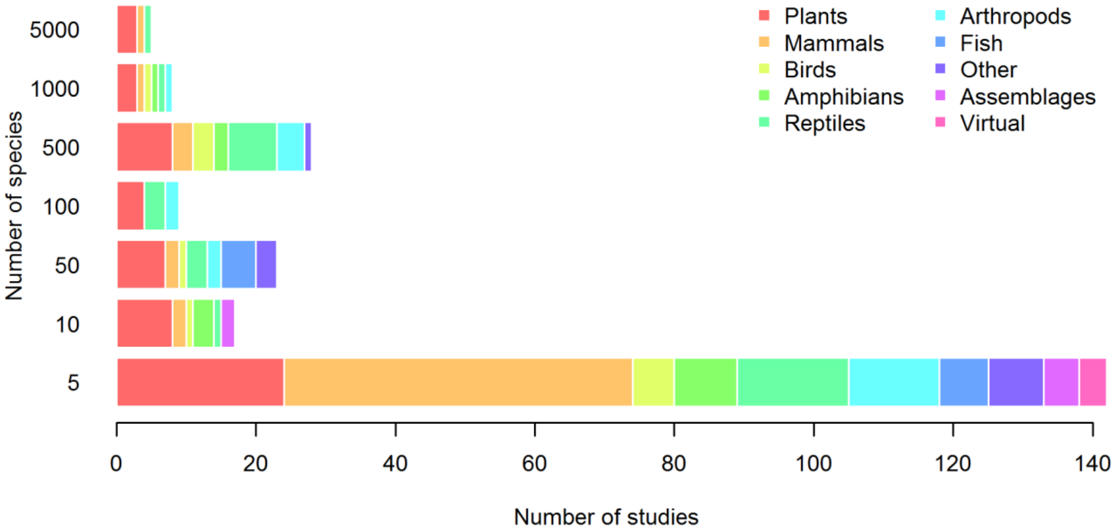


Fig. S1.1. Classification of 400 randomly sampled papers applying SDMs to biodiversity assessments according to the number and taxonomic group of species modeled.

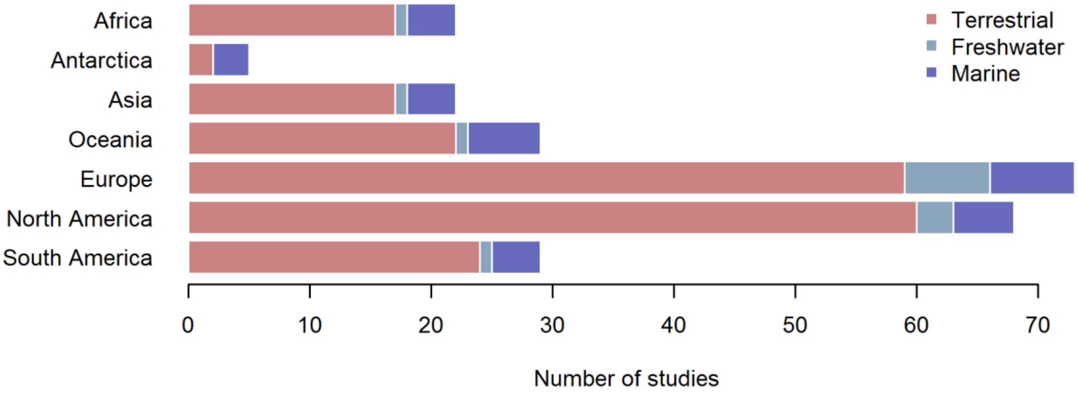
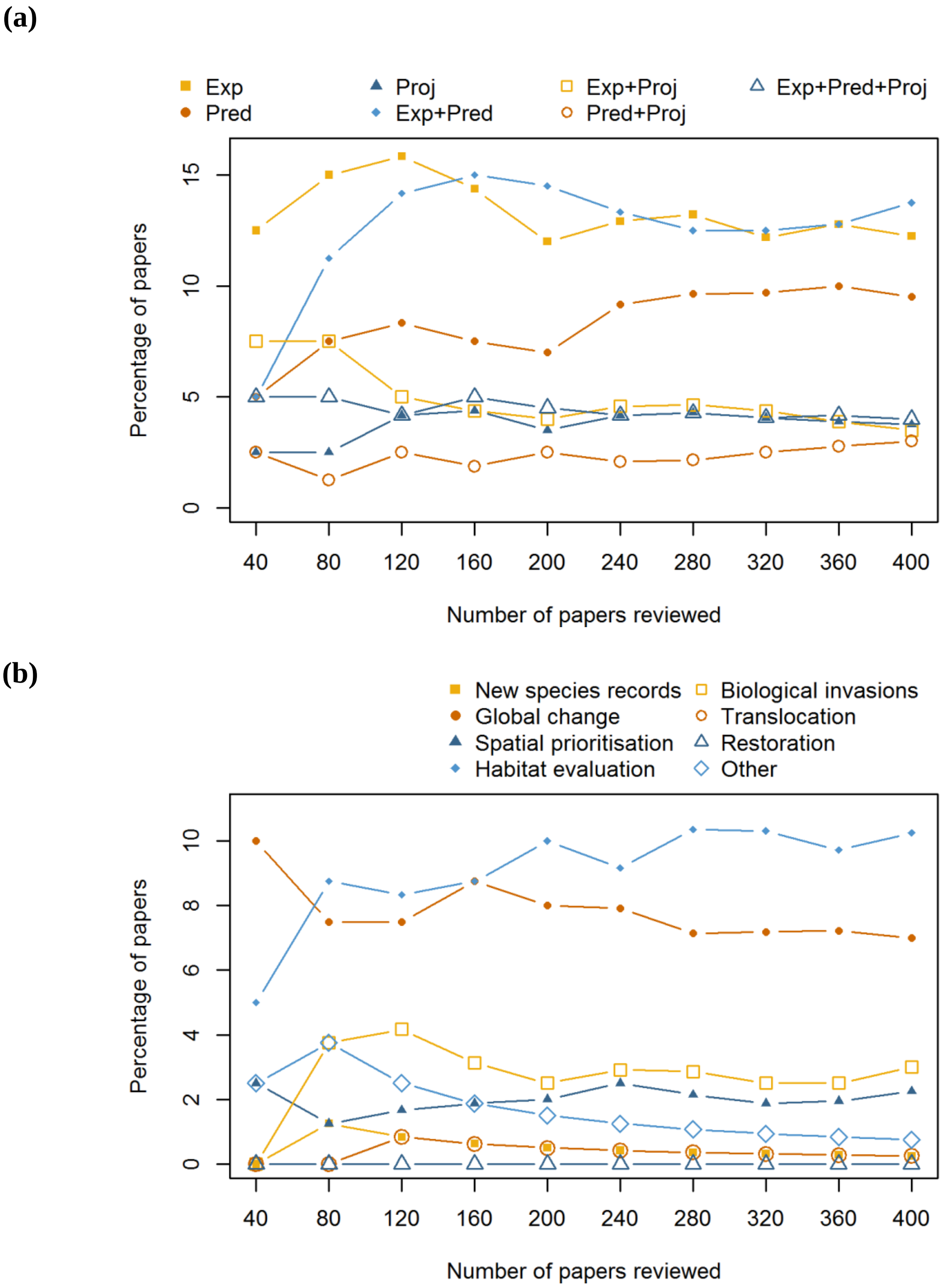
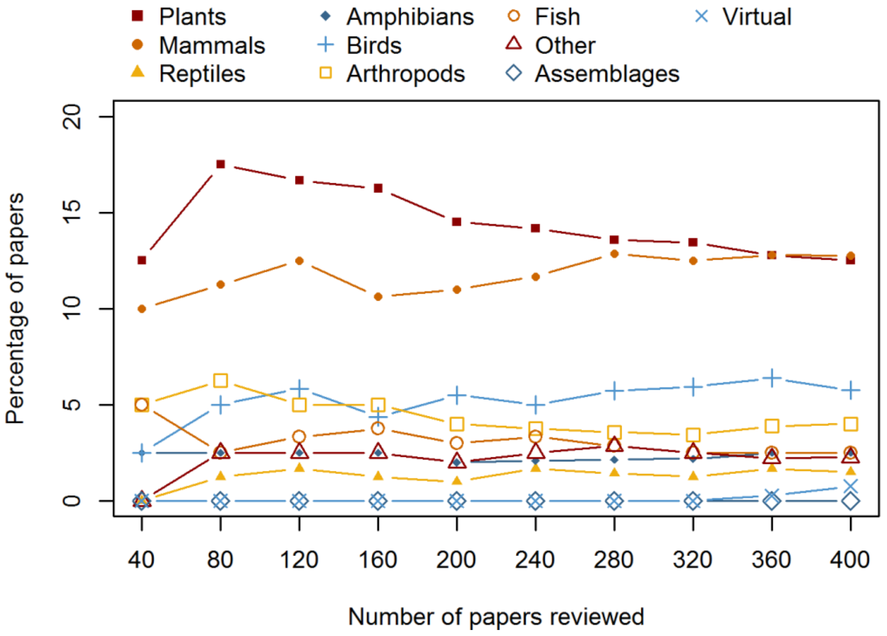


Fig. S1.2. Classification of 400 randomly sampled papers applying SDMs to biodiversity assessments according to the continent and ecological realm of focus.

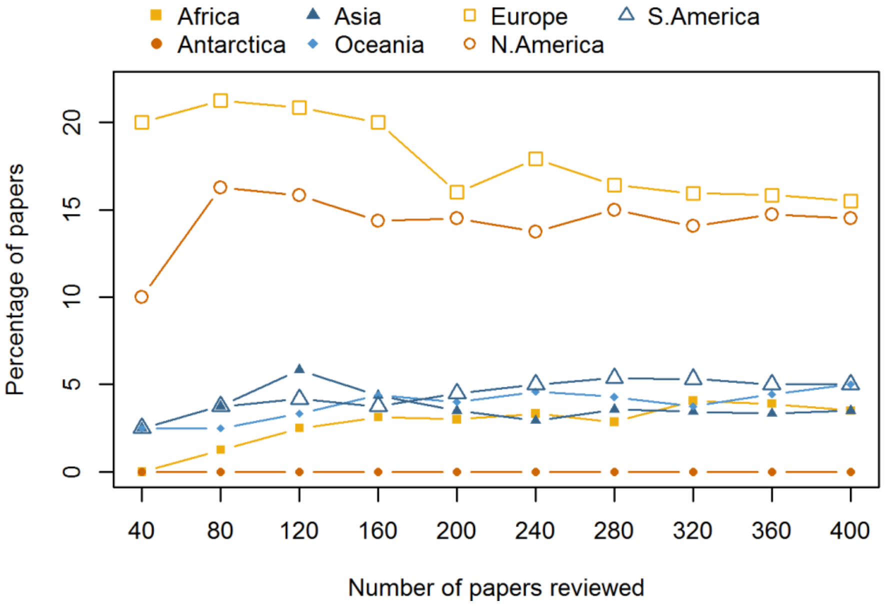

(a)
Exp
Pred
Proj
Exp+Pred
Exp+Proj
Pred+Proj
Exp+Pred+Proj
Percentage of papers
0
5
10
15
40
80
120
160
200
240
280
320
360
400
Number of papers reviewed
(b)
New species records
Global change
Spatial prioritisation
Habitat evaluation
Biological invasions
Translocation
Restoration
Other
Percentage of papers
0
2
4
6
8
10
40
80
120
160
200
240
280
320
360
400
Number of papers reviewed

(c)



Plants
Mammals
Reptiles
Amphibians
Birds
Arthropods
Fish
Other
Assemblages
Virtual
Percentage of papers
0
5
10
15
20
40
80
120
160
200
240
280
320
360
400
Number of papers reviewed

(d)



Africa
Antarctica
Asia
Oceania
Europe
N.America
S.America
Percentage of papers
0
5
10
15
20
40
80
120
160
200
240
280
320
360
400
Number of papers reviewed

(e)

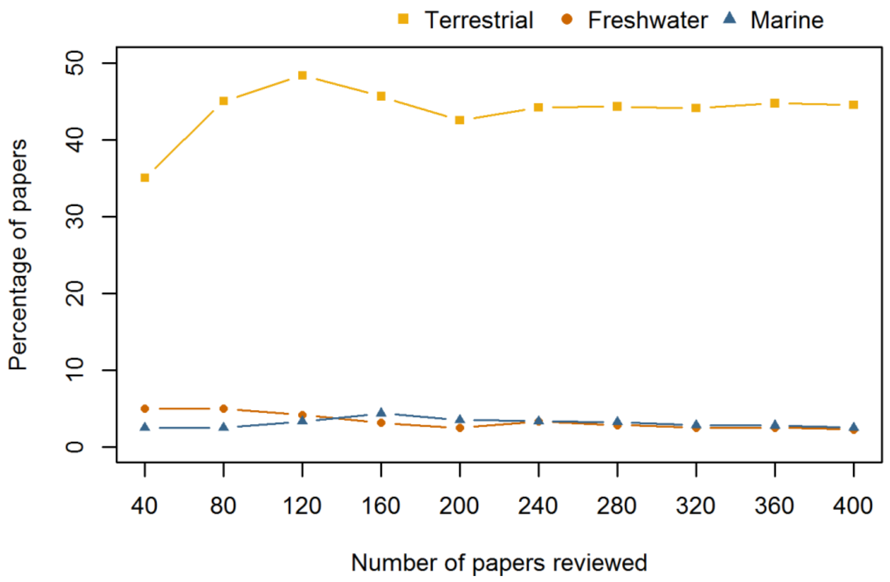


Fig. S1.3. Accumulated percentage of papers reviewed falling in different classes as the size of the random sample is increased. The different classifications are: (**a**) the purpose of the species distribution model (table S1.2), (**b**) the intended conservation application (table S1.3), (**c**) the taxonomic group modelled, and the (**d**) continent or (**e**) ecological realm where the species is present.

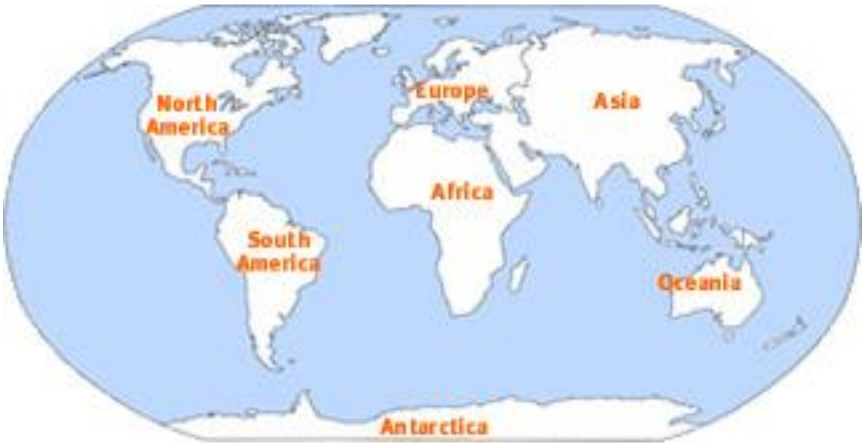
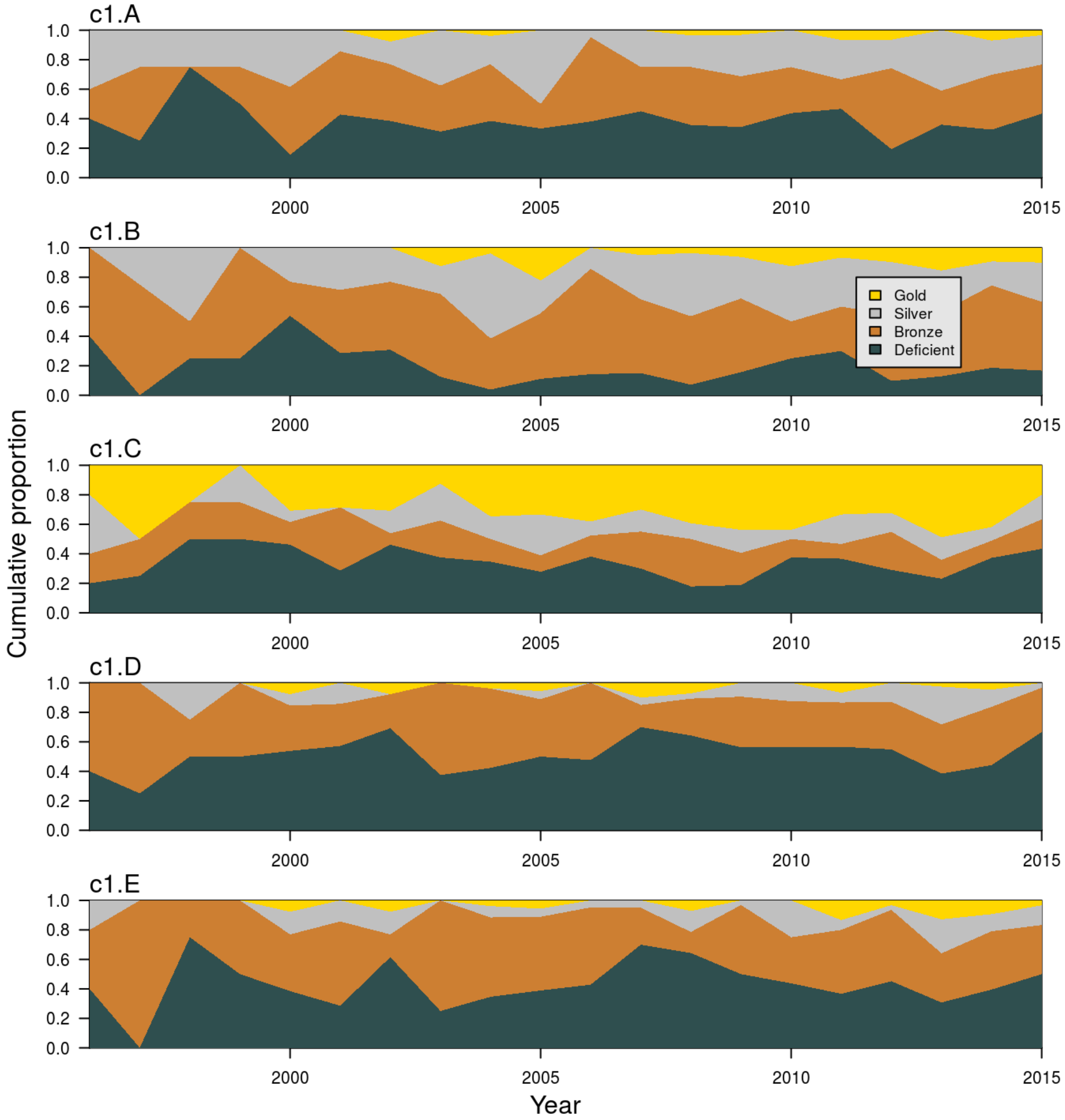
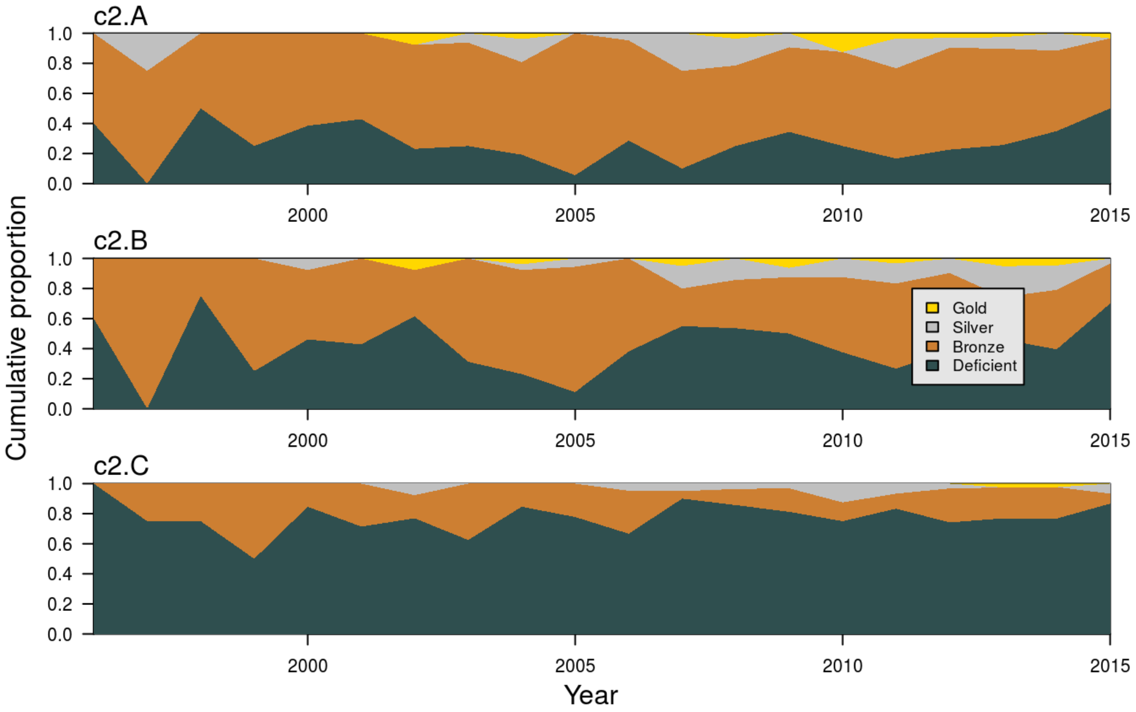


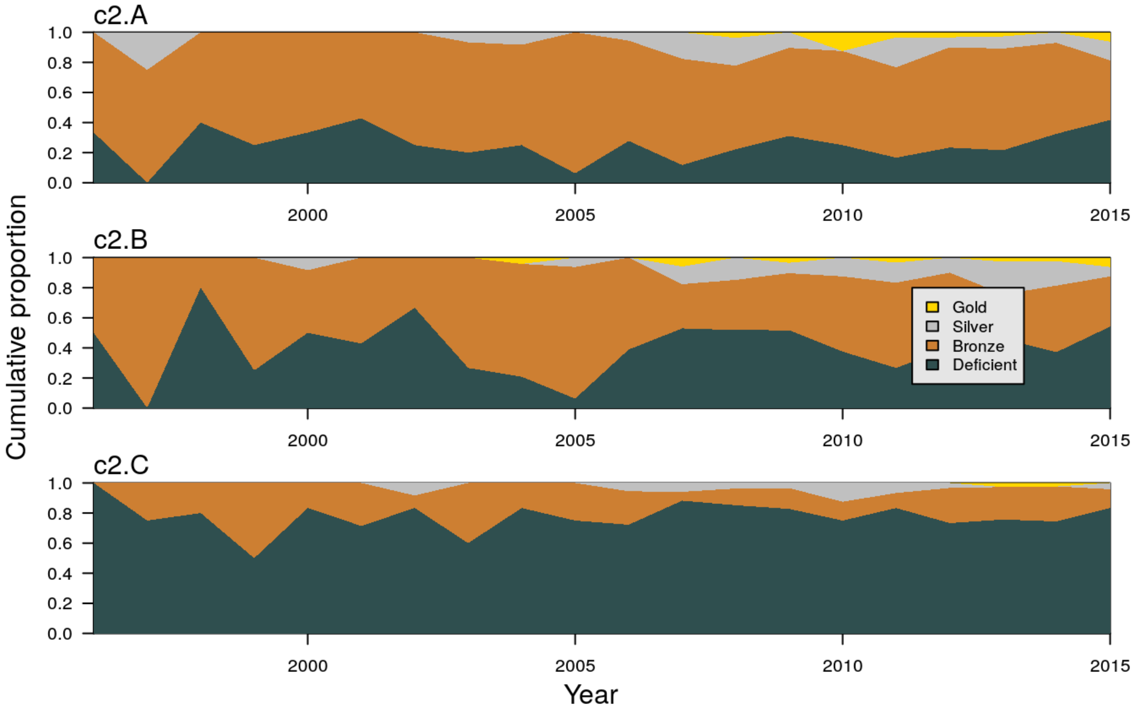
Fig. S1.4. The continents used to classify papers applying SDMs to biodiversity assessments. Map reproduced from image released into the public domain by http://www.blatantworld.com/, licensed under Creative Commons License type Attribution 2.0 Generic.

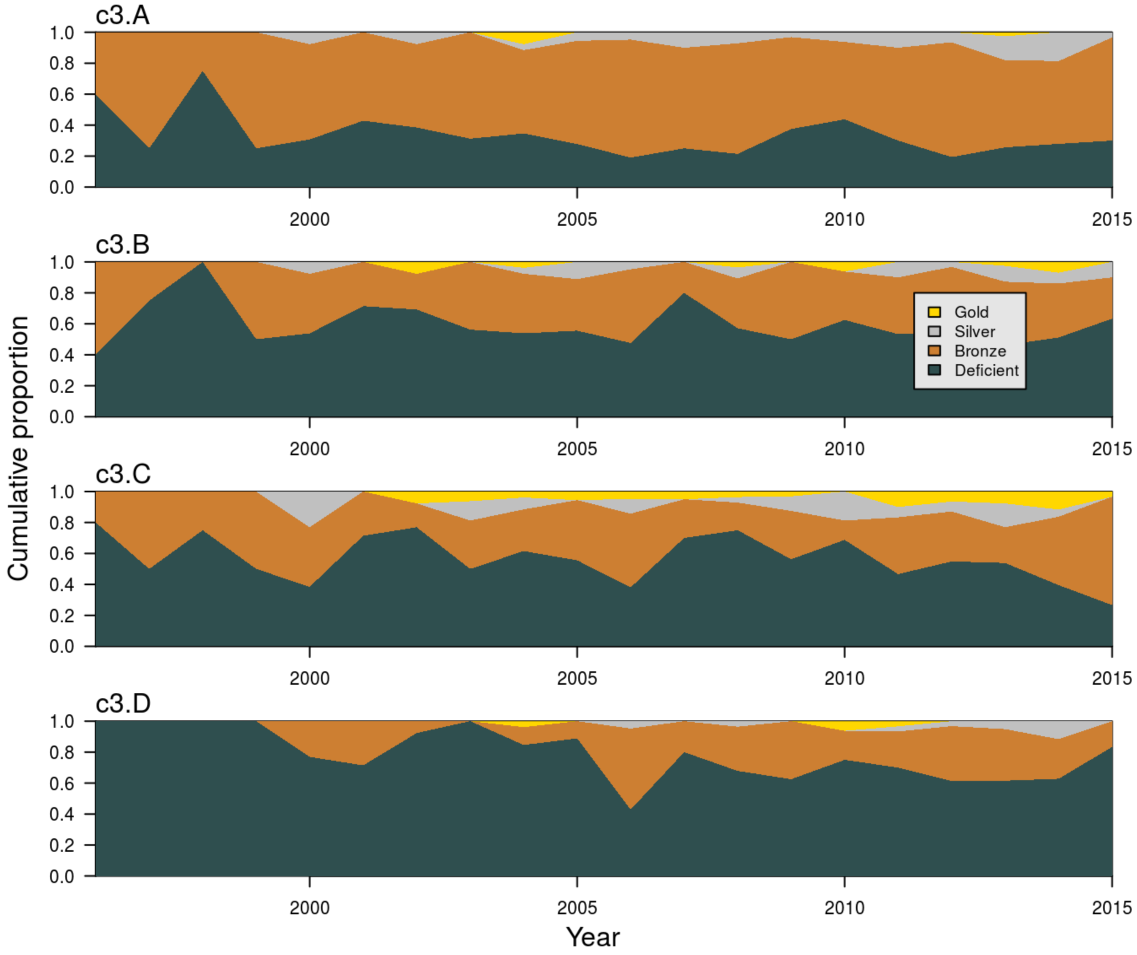
RESPONSE VARIABLE



PREDICTOR VARIABLES






c3.A
c3.B
c3.C
c3.D
1.0
0.8
0.6
0.4
0.2
0.0
2000
2005
2010
2015
Cumulative proportion
Year
Gold
Silver
Bronze
Deficient

MODEL EVALUATION

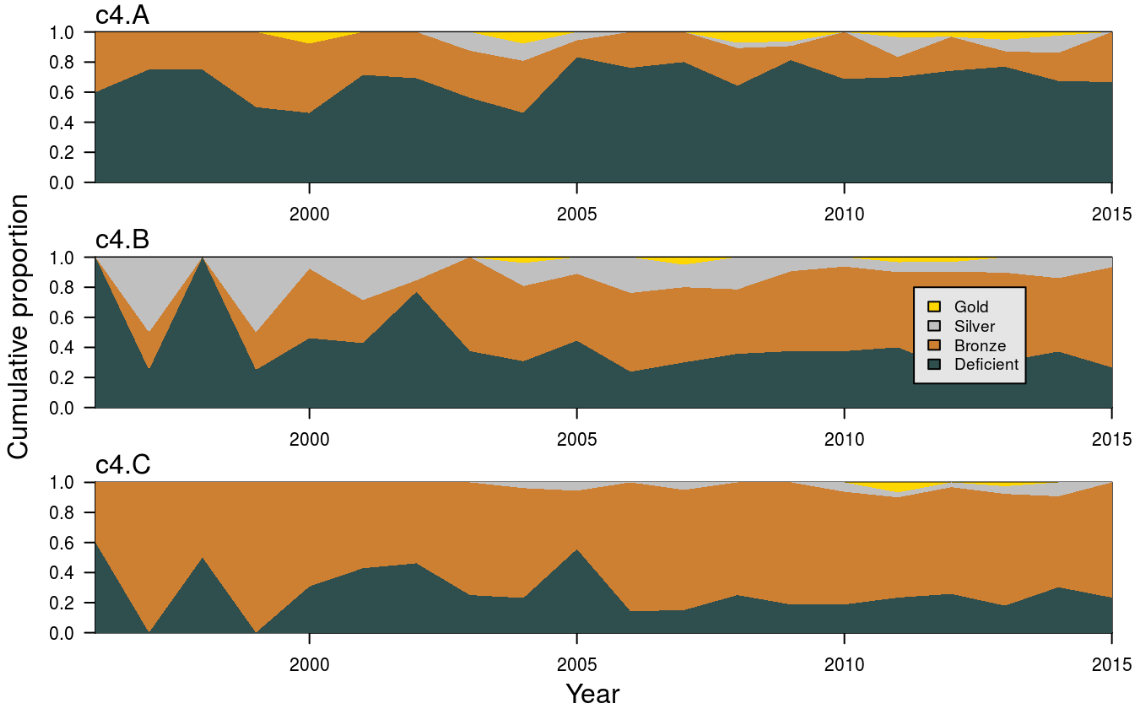


Fig. S1.5. Frequencies of scores of different categories of issues assessed. 1.A sampling response variables; 1.B identification of taxa; 1.C spatial accuracy of response variable; 1.D environmental extent across which response variable is sampled; 1.E geographical extent across which response variable is sampled; 2.A selection of candidate variables; 2.B spatial and temporal resolution of predictor variables; 2.C uncertainty in predictor variables; 3.A model complexity; 3.B treatment of bias and noise in response variables; 3.C treatment of collinearity; 3.D dealing with modelling and parameter uncertainty; 4.A evaluation of model assumptions; 4.B evaluation of model outputs; 4.C measures of model performance.

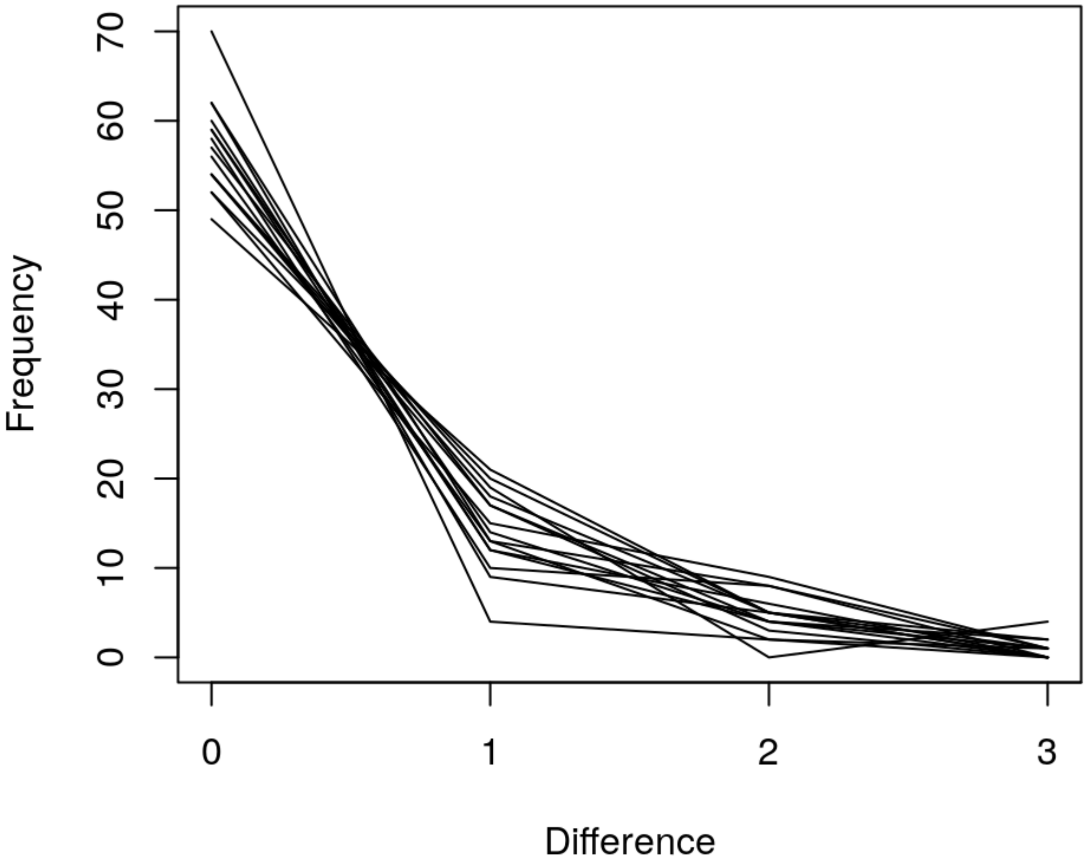
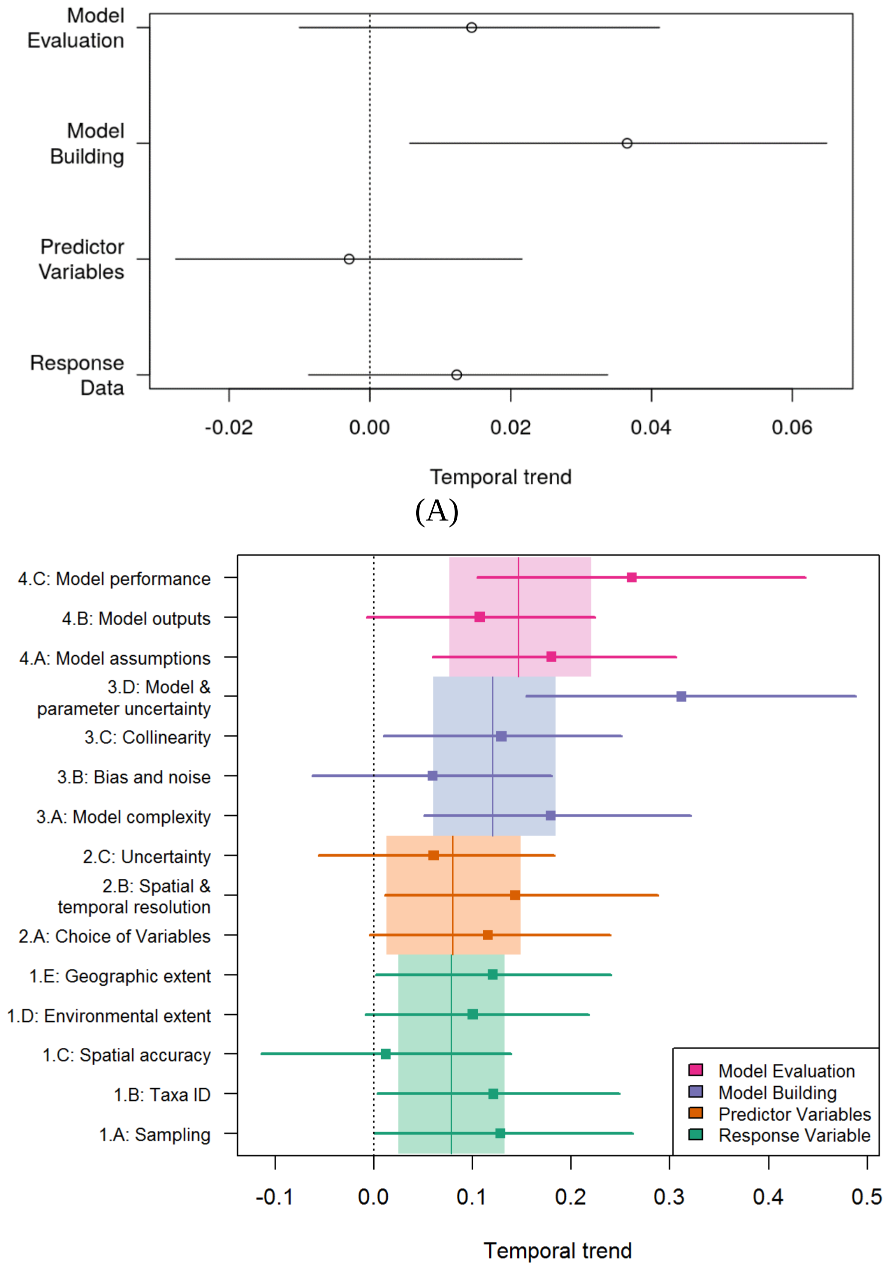


Fig. S1.6. Differences between scores obtained in the first assessment of the studies and the second independent reevaluation by a different assessor. Summary across all aspects and criteria judged with 0 in the x-axis representing no change between evaluation and independent re-evaluation and 3 representing the greatest possible difference. See text S3 for explanation.



(A)

(B)

Fig. S1.7. Changes in species distribution modeling standards over time (1995–2015). (**A**) Data aggregated for each one of the four critical aspects of data and models examined. (**B**) Data aggregated for each one of the 15 issues and restricted to the studies with scores at or above 90%-ile for each 5-years time blocks (equivalent analysis for all data, provided in Figure 4). The diagrams show the results of ordinal regression using 'Year' as a continuous variable and the four key aspects of modelling as effects (including an interaction). Values near zero on the x-axis represent no change in standards over time, positive values indicate improvement, and bars are 95% credible intervals.

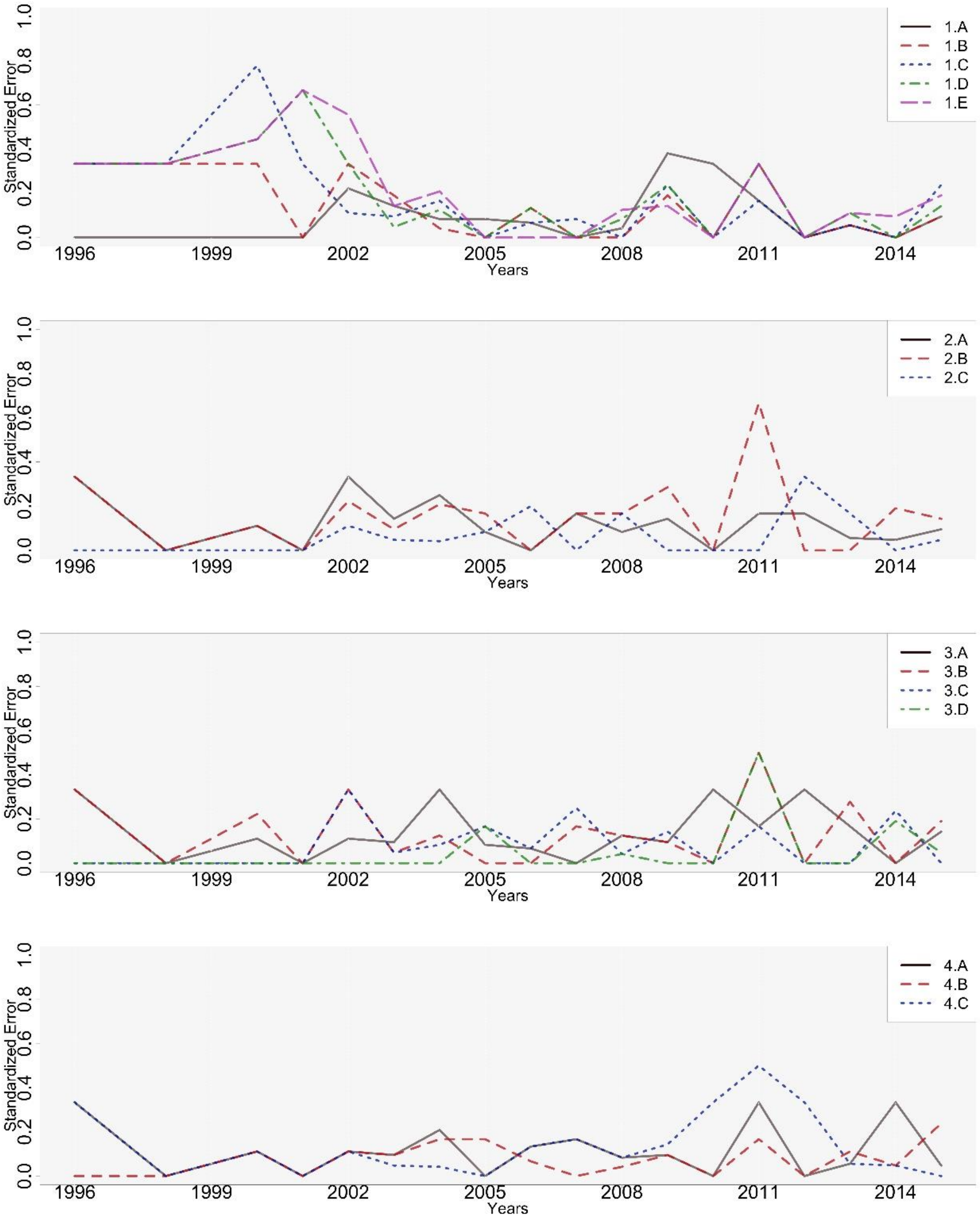


Fig. S1.8. Magnitude of standard deviations (0, no error; 1, maximum error) between first and second independent scoring of the studies over annual steps for each aspect and issue judged. See text S3 for explanation.

Supplementary tables

Table S1.1. Search terms used to select papers for the literature characterization.

Topic	Search terms
Species distribution models	TS=(“species distribut*” OR “habitat distribut*" OR "climat* envelope" OR bioclimat* OR "habitat suitab*" OR niche OR "resource selection" OR SDM OR ENM OR BEM OR BCM OR HSM OR RSF) AND (model*)
Biodiversity context	TS=(biolog* OR nature OR species OR habitat OR environment* OR ecosystem OR ecology OR wildlife OR biodivers*)
Time period	PY=(1996-2015)
Journals	SO=(ANNU REV ECOL EVOL S OR TRENDS ECOL EVOL OR ECOL LETT OR ECOL MONOGR OR FRONT ECOL ENVIRON OR ISME J OR GLOBAL CHANGE BIOL OR METHODS ECOL EVOL OR GLOBAL ECOL BIOGEOGR OR CONSERV LETT OR MOL ECOL OR J ECOL OR ECOLOGY OR J APPL ECOL OR P ROY SOC B-BIOL SCI OR ECOGRAPHY OR ECOL APPL OR DIVERS DISTRIB OR J ANIM ECOL OR FUNCT ECOL OR EVOLUTION OR CONSERV BIOL OR J BIOGEOGR OR B AM MUS NAT HIST OR WILDLIFE MONOGR OR MOL ECOL RESOUR OR AM NAT OR ADV ECOL RES OR BIOL CONSERV OR BIOGEOSCIENCES OR ECOSYSTEMS OR PERSPECT PLANT ECOL OR HEREDITY OR ECOL SOC OR AGR ECOSYST ENVIRON OR ECOL ECON OR OIKOS OR LANDSCAPE ECOL OR J VEG SCI OR BIOL LETTERS OR LANDSCAPE URBAN PLAN OR OECOLOGIA OR J EVOLUTION BIOL OR MICROB ECOL OR BEHAV ECOL OR ECOL ENG OR ANIM CONSERV OR MAR ECOL PROG SER OR ECOHYDROLOGY OR ENVIRON CONSERV OR PALEOBIOLOGY OR ECOTOXICOLOGY OR BEHAV ECOL SOCIOBIOL OR FUNGAL ECOL OR BIOL INVASIONS OR ECOL COMPLEX OR J PLANT ECOL OR J CHEM ECOL OR ECOL MODEL OR BASIC APPL ECOL OR BIODIVERS CONSERV OR ECOSPHERE OR EVOL ECOL OR AQUAT MICROB ECOL ORBIOTROPICA OR APPL VEG SCI OR J EXP MAR BIOL ECOL OR ECOL EVOL OR EUR J SOIL BIOL OR RESTOR ECOL OR ECOL INFORM OR AM MUS NOVIT OR ORYX OR RANGELAND ECOL MANAG OR J ARID ENVIRON OR J WILDLIFE MANAGE OR PEDOBIOLOGIA OR SYST BIODIVERS OR THEOR ECOL-NETH OR FRESHW SCI OR POPUL ECOL OR WETLANDS OR AQUAT ECOL OR ACTA OECOL OR PLANT ECOL OR POLAR RES OR CONSERV GENET OR J SOIL WATER CONSERV OR AUSTRAL ECOL OR J NAT CONSERV OR POLAR BIOL OR THEOR POPUL BIOL OR ECOL RES OR WILDLIFE RES OR CHEMOECOLOGY OR FLORA OR COMMUNITY ECOL OR EUR J WILDLIFE RES OR AVIAN CONSERV ECOL OR NEW ZEAL J ECOL OR ECOSCIENCE OR B PEABODY MUS NAT HI OR ENVIRON BIOL FISH OR MAR BIOL RES OR J TROP ECOL OR NAT CONSERVACAO OR WILDLIFE BIOL OR TROP CONSERV SCI OR POL POLAR RES OR RANGELAND J OR ANN ZOOL FENN OR CHEM ECOL OR BIOCHEM SYST ECOL OR TROP ECOL OR EVOL ECOL RES OR PLANT SPEC BIOL OR LANDSC ECOL ENG OR S AFR J WILDL RES OR POLAR SCI OR J NAT HIST OR CONSERV GENET RESOUR OR J FISH WILDL MANAG OR REV CHIL HIST NAT OR NAT AREA J OR AFR J ECOL OR AM MIDL NAT OR P ACAD NAT SCI PHILA OR COMPOST SCI UTIL OR AFR J RANGE FOR SCI OR NORTHWEST SCI OR APPL ECOL ENV RES OR POLAR REC OR EKOLOJI OR POL J ECOL OR ISR J ECOL EVOL OR J FRESHWATER ECOL OR REV MEX BIODIVERS OR CARIBB J SCI OR VIE MILIEU OR ECOTROPICA OR NORTHEAST NAT OR WEST N AM NATURALIST OR RUSS J ECOL+ OR PACHYDERM OR ECO MONT OR SOUTHEAST NAT OR INTERCIENCIA OR REV ECOL-TERRE VIE OR SOUTHWEST NAT OR CONTEMP PROBL ECOL+ OR NAT HIST OR ANIM BIODIV CONSERV OR AQUAT INVASIONS OR BIOTA NEOTROP OR BMC ECOL OR FIRE ECOL OR J BIOL DYNAM OR KOEDOE OR MAR BIODIVERS OR P LINN SOC N S W OR URBAN ECOSYST)

Table S1.2. Classification of the purpose for which SDMs are used.

Classes	Definition
Explanation	Investigate a species' (causal) relationship with the environment.
Prediction	Map species' potential distributions within the same time period and geographic region as the data used to construct SDMs.
Projection	Project species distribution predictions into a different time period or location from the data used to construct SDMs. Also called 'transferring'.

Table S1.3. Classification of the conservation applications of SDMs.

Conservation application	
New species records	Spatially identifies areas in which the species is not currently recorded but might be found, and which should be surveyed.
Global change	Predicts spatial locations of areas that will change in suitability due to climate or land-use change.
Spatial prioritisation	Spatially identifies areas in which conservation would be valuable, or calculates a metric of conservation value of specific areas.
Habitat evaluation	Quantifies ability of landscape to support existing populations, or population decline in the landscape. Includes population viability analyses. Does not make spatially explicit recommendations about habitat management.
Biological invasions	Spatially identifies areas that could be threatened by, facilitate, or prevent biological invasions or the transmission of disease.
Translocation	Spatially identifies areas suitable for translocated populations, or calculates suitability of specific areas.
Restoration	Spatially identifies areas that would be appropriate for restoration, or measures the efficacy of restoring specific areas.

Table S2.1. Guidelines—Response variable.

Issue	Explanation	Prediction	Projection	Standard
1A. Sampling of response variables	Same as prediction or projection, depending on whether desired explanation is local or global respectively.	Sampled via systematically designed surveys demonstrated to encompass the major environmental gradients occupied by the taxon, and spatial extent of the taxon's occurrence, within the study area. Includes estimates of population demographic parameters (to identify self-sustaining populations), and taxon detectability. Information available on intensity of sampling at each site, and used to ensure sampling is unbiased.	Sampled via systematically designed surveys demonstrated to encompass the major environmental gradients occupied by the taxon, and spatial extent of the taxon's occurrence. Includes estimates of population demographic parameters (to identify self-sustaining populations) and taxon detectability. Information available on intensity of sampling at each site, and used to ensure sampling is unbiased.	Gold
	Same as prediction or projection, depending on whether desired explanation is local or global respectively.	Sampled via systematically designed surveys that encompass the major environmental gradients occupied by the taxon, and spatial extent of the taxon's occurrence, within the study area. Information on intensity of sampling at each site used to conduct post-hoc resampling/weighting to reduce bias; see box 3B.	Sampled via systematically designed surveys that encompass the major environmental gradients occupied by the taxon, and spatial extent of the taxon's occurrence. Information on intensity of sampling at each site used to conduct post-hoc resampling/weighting to reduce bias; see box 3B.	Silver
	Sampled via non-systematically designed surveys, with information on intensity of sampling at each site used to conduct post-hoc resampling/weighting to reduce bias; see box 3B). OR Sampled via non-systematically designed surveys, without information on intensity of sampling at each site, but post-hoc processing undertaken to reduce bias and yield geographically and environmentally representative samples; see box 3B).			Bronze
	Sampled via non-systematically designed surveys. No post-hoc resampling / weighting / processing to reduce bias or yield geographically and environmentally representative samples.			Deficient
1B. Identification of taxa (if species	ID provided by experts, based on multiple lines of evidence, which can be examined.			Gold
	ID provided by experts, based on a single line of evidence, which can be examined.			Silver
	ID based on heterogeneous sources. Records used without being checked by taxonomic experts but after being critically "cleaned" to			Bronze

Issue	Explanation	Prediction	Projection	Standard
occurrence used as response variable)	remove unreasonable records.			
	ID based on heterogeneous sources. Records used without being checked by taxonomic experts and without being critically “cleaned” by others to remove unreasonable records.			Deficient

Issue	Explanation	Prediction	Projection	Standard
1C. Spatial accuracy of response variable	Spatial accuracy of all records sufficiently high relative to spatial resolution of predictor variables such that all points are known to fall within the location to which they are assigned. The spatial accuracy matches the spatial resolution of predictor variables as defined in section 2B.			Gold
	Spatial accuracy of all records known and variable across records, so that some points might fall outside the location to which they are assigned. These potential locational errors integrated into formal uncertainty analysis (see 3D), and/or steps taken (and documented) to remove records with locational errors.			Silver
	Spatial accuracy not known or inconsistently quantified, but steps taken (and documented) to remove unreasonable records.			Bronze
	Spatial accuracy not known or inconsistently quantified, and no steps taken to remove unreasonable records.			Deficient
1D. Environmental extent across which response variable is sampled	Same as prediction or projection, depending on whether desired explanation is local or global respectively.	Multiple lines of evidence (in addition to occurrence data used to train SDMs) demonstrate that data cover the range of the taxon's environmental tolerances within the study area.	Multiple lines of evidence (in addition to occurrence data used to train SDMs) demonstrate that data cover the entire range of the taxon's environmental tolerances and that no evolutionary changes in species environmental tolerances have occurred in the projection space.	Gold
	Same as prediction or projection, depending on whether desired explanation is local or global respectively.	A single line of evidence (in addition to occurrence data used to train SDMs) demonstrates that data cover the entire environmental range of the study region.	A single line of evidence (in addition to occurrence data used to train SDMs) demonstrates that data cover the entire environmental extent of the known distribution of the taxon.	Silver
	Models fitted with the best available data on the known geographical extent of the taxon, but without evidence that the taxon's environmental tolerances are covered. Steps taken and documented to avoid impact of incomplete distribution data on results.	Steps are taken to avoid or flag extrapolation to conditions outside the environmental extent used to train the models.	Models fitted with the best available data on the known geographical extent of the taxon, but without evidence that the taxon's environmental tolerances are covered. Steps taken to avoid or flag extrapolation to conditions outside the extent of each predictor variable used to train the models.	Bronze
	No evidence provided that data cover the entire environmental range of the taxon.	No evidence provided that data cover the entire environmental range of the study region.	No evidence provided that data cover the entire environmental range of the study region. No steps taken to avoid or flag extrapolation to conditions outside the extent of each predictor used to train the models.	Deficient

Issue	Explanation	Prediction	Projection	Standard
1E. Geographic extent across which response variable is sampled (includes occurrence data and absence, pseudo-absence, or background data	Same as prediction or projection, depending on whether desired explanation is local or global respectively.	Samples restricted to and inclusive of all regions of the study area that are suitable for the taxa to establish populations, and that are accessible to the taxon (as demonstrated by multiple lines of evidence).	Samples restricted to and inclusive of all regions that are suitable for the taxa to establish populations and that are accessible to the taxon (as demonstrated by multiple lines of evidence).	Gold
	Same as prediction or projection, depending on whether desired explanation is local or global respectively.	Samples restricted to and inclusive of all regions of the study area that contain the full historical and current range of the focal taxon.	Samples restricted to and inclusive of all regions that contain the full historical and current range of the focal taxon (as demonstrated by a single line of evidence).	Silver
	Same as prediction or projection, depending on whether desired explanation is local or global respectively.	Samples derived only from regions within the study area reasonably justified to contain the full current range of the focal taxon.	Samples derived only from regions reasonably justified to contain the full current range of the focal taxon.	Bronze
	No justification of regions from which samples drawn, or samples derive from regions outside those reasonably deemed accessible to the taxon.			Deficient

Table S2.2. Guidelines—Predictor variables.

Issue	Explanation	Prediction	Projection	Standard
2A. Selection of candidate variables	Candidates include all proximal variables that multiple lines of evidence (in addition to occurrence data used to train SDMs) that can be shown to have a measurable effect on the taxon's distribution at the spatial scale examined. This must include, whenever relevant, a full range of environmental and biotic variables.			Gold
	Same as prediction or projection, depending on whether desired explanation is local or global respectively.	Candidates include proximal and/or distal variables that a single line of evidence (in addition to occurrence data used to train SDMs) shows to have a measurable association with the taxon at the spatial scale examined. This should include, whenever relevant, a range of environmental and biotic variables.	Candidates include proximal variables that a single line of evidence (in addition to occurrence data used to train SDMs) shows to have a measurable effect on the taxon's distribution at the spatial scale examined. This should include, whenever relevant, a range of environmental and biotic variables.	Silver
	Same as prediction or projection, depending on whether desired explanation is local or global respectively.	Candidates include proximal or distal variables theoretically justified as having an association with the taxon's distribution at the spatial scale examined.	Candidates include observationally, statistically or theoretically justified proximal and/or distal variables that have a measurable association with the taxon's distribution at the spatial scale examined. This should include, whenever possible, a range of environmental and biotic variables.	Bronze
	Same as prediction or projection, depending on whether desired explanation is local or global respectively.	No ecological justification of variable choice.	No ecological justification of variable choice and/or distal variables used without strong justification.	Deficient
2B. Spatial and temporal resolution of predictor variables	Variables directly measured at the temporal and spatial resolution at which multiple lines of evidence (in addition to occurrence data used to train SDMs) demonstrate that the taxon responds.			Gold
	Variables interpolated at the temporal and spatial resolution at which at least one line of evidence (in addition to occurrence data used to train SDMs) demonstrate the taxon responds.			Silver
	Variables interpolated at a resolution theoretically justified for the taxon.			Bronze
	Variables interpolated at a spatial and temporal resolution to which the taxon does not respond and/or without theoretical justification of resolution.			Deficient

Issue	Explanation	Prediction	Projection	Standard
2C. Uncertainty in predictor variables (both under current and projected conditions)	All sources of uncertainty in the predictors and their effects on model results quantified, mapped, and interpreted.			Gold
	Some of the perceived most important sources of uncertainty in the predictors (e.g. errors in geo-registration, measurement, interpolation) quantified and mapped.			Silver
	Possible sources of uncertainty in the predictors (e.g. errors in geo-registration, measurement, interpolation) and the effects these could have on the model acknowledged, and consequences for interpretation of the results discussed.			Bronze
	No consideration of uncertainty.			Deficient

Table S2.3. Guidelines—Model building.

Issue	Explanation	Prediction	Projection	Standard
3A. Model complexity	Same as prediction or projection, depending on whether desired explanation is local or global respectively.	Same as projection without the necessity of using independent data.	The optimal level of complexity is decided by constructing models using an appropriate method to deal with model complexity, performing comparison with multiple lines of independent data (see table 4B).	Gold
	Same as prediction or projection, depending on whether desired explanation is local or global respectively.	Same as projection without the necessity of using independent data.	The optimal level of complexity is decided by constructing models using an appropriate method to deal with model complexity, performing cross-validation or comparison with a single line of independent data.	Silver
	Broadly agreed rules of thumb are followed and/or the optimal level of complexity is decided using justified methods without independent data.			Bronze
	Complexity is not considered, or inappropriate methods are used to deal with it.			Deficient
3B. Treatment of bias and noise in response variables	Demonstrated that there are no geographical and environmental biases in response data. OR Model fully corrected for bias in response data, tested by performing comparison with independent data.			Gold
	Model corrected for major biases in response data, tested by performing internal cross-validation.			Silver
	Bias, and the effects these could have on the model and results, acknowledged and described.			Bronze
	No consideration of biases.			Deficient

Issue	Explanation	Prediction	Projection	Standard
3C. Treatment of collinearity	Demonstrated that there is no collinearity in data. OR Model construction is informed by a full mechanistic understanding of interactions among predictor variables so that the model is insensitive to collinearity.			Gold
	Fitting techniques used are known to be insensitive to collinearity. OR Demonstrated that the results are robust to changes in collinearity between predictor variables, including non-analogue combinations of predictor variables.			Silver
	Approximate methods are applied to deal with collinearity. OR Collinearity is acknowledged and described, as are the effects the known collinearity could have on the results.			Bronze
	Models use collinear variables and a fitting technique sensitive to collinearity without acknowledging the effects on the results,			Deficient
3D. Dealing with modelling and parameter uncertainty	Uncertainty arising from different modelling techniques, response data, and predictor variables is comprehensively characterized. Results are obtained from several SDM techniques that are representative of all appropriate current distribution modelling techniques in order to characterize model uncertainty (sometimes called ensemble modelling). Uncertainty is fully propagated through the modelling process in order to quantify, map, and interpret uncertainty in results. Biases arising from similarities among structures of model classes, and the effects these could have on results are discussed.			Gold
	Major suspected model and data uncertainties are characterized and: 1) known uncertainties are propagated through the model; or 2) the range of predictions built using different scenarios (including parameter and model technique) are quantified and mapped, and sensitivity analysis is conducted. Results are obtained from several SDM techniques that are representative of all appropriate current distribution modelling techniques in order to characterize model uncertainty. Biases arising from similarities among structures of model classes are quantified, accounted for, and discussed.			Silver
	Results are obtained from multiple SDM techniques but that are not representative of all appropriate current distribution modelling techniques. The effect of major suspected model uncertainties on the projections is quantified. Major suspected sources of data uncertainties are acknowledged, and their consequences for interpretation of the results are discussed.			Bronze
	Uncertainty is not dealt with (i.e., a single SDM technique with one set of parameters is used).			Deficient

Table S2.4. Guidelines—Model evaluation.

Issue	Explanation	Prediction	Projection	Standard
4A. Evaluation of model assumptions	Demonstrated lack of violation of, or robustness to, assumptions relevant for technique being used.			Gold
	Theoretically justified lack of violation of, or expected robustness to, assumptions of technique being used.			Silver
	Violation of major assumptions of technique being used characterized, and their consequences for interpretation of results discussed.			Bronze
	No check for violation of statistical assumptions.			Deficient
4B. Evaluation of model outputs	Same as prediction or projection, depending on whether desired explanation is local or global respectively.	Evaluated against multiple datasets that are statistically independent from the data used to train the models, but not necessarily from an independent location or time period.	Evaluated against multiple and diverse independent evaluation datasets, and/or corroboration with experimental testing.	Gold
	Same as prediction or projection, depending on whether desired explanation is local or global respectively.	Evaluated against data obtained by geographically structured sub-sampling of the training data.	Evaluated against at least one independent evaluation dataset.	Silver
	Same as prediction or projection, depending on whether desired explanation is local or global respectively.	Evaluated with non-independent data. Re-substitution used to estimate over-fitting. N – 1 Jackknife acceptable for very small sample sizes.	Evaluated with non-independent data obtained by sub-sampling the training data, with repetition. N – 1 Jackknife acceptable for very small sample sizes.	Bronze
	No evaluation at all or re-substitution alone.			Deficient

Issue	Explanation	Prediction	Projection	Standard
4C. Measures of model performance	Same as prediction.	Same as projection, but no characterization of temporal errors.	Measures of performance exhaustively summarize goodness of fit and discrimination. AND Patterns of uncertainty comprehensively characterized (spatial, temporal and environmental; see also 3E).	Gold
	Measures of performance summarize goodness of fit and calibration AND Major patterns of uncertainty comprehensively characterized (spatial, temporal and/or environmental; see also 3E)			Silver
	One or more major aspects of model performance measured and summarized.			Bronze
	No, cursory, or inappropriate measures of model performance.			Deficient

Table S3.1. Search terms used to select papers using SDMs for biodiversity assessments, for the purpose of scoring according to the guidelines.

Topic	Search terms
Species distribution models	TS=(“species distribut*” OR “habitat distribut*" OR "climat* envelope" OR bioclimat* OR "habitat suitab*" OR niche OR "resource selection" OR SDM OR ENM OR BEM OR BCM OR HSM OR RSF) AND TS=(model*)
Biodiversity context	TS=(biolog* OR nature OR species OR habitat OR environment* OR ecosystem OR ecology OR wildlife OR biodivers*)
Time period	PY=(1996-2015)
Journals	SO=(ANNU REV ECOL EVOL S OR TRENDS ECOL EVOL OR ECOL LETT OR ECOL MONOGR OR FRONT ECOL ENVIRON OR ISME J OR GLOBAL CHANGE BIOL OR METHODS ECOL EVOL OR GLOBAL ECOL BIOGEOGR OR CONSERV LETT OR MOL ECOL OR J ECOL OR ECOLOGY OR J APPL ECOL OR P ROY SOC B-BIOL SCI OR ECOGRAPHY OR ECOL APPL OR DIVERS DISTRIB OR J ANIM ECOL OR FUNCT ECOL OR EVOLUTION OR CONSERV BIOL OR J BIOGEOGR OR B AM MUS NAT HIST OR WILDLIFE MONOGR OR MOL ECOL RESOUR OR AM NAT OR ADV ECOL RES OR BIOL CONSERV OR BIOGEOSCIENCES OR ECOSYSTEMS OR PERSPECT PLANT ECOL OR HEREDITY OR ECOL SOC OR AGR ECOSYST ENVIRON OR ECOL ECON OR OIKOS OR LANDSCAPE ECOL OR J VEG SCI OR BIOL LETTERS OR LANDSCAPE URBAN PLAN OR OECOLOGIA OR J EVOLUTION BIOL OR MICROB ECOL OR BEHAV ECOL OR ECOL ENG OR ANIM CONSERV OR MAR ECOL PROG SER OR ECOHYDROLOGY OR ENVIRON CONSERV OR PALEOBIOLOGY OR ECOTOXICOLOGY OR BEHAV ECOL SOCIOBIOL OR FUNGAL ECOL OR BIOL INVASIONS OR ECOL COMPLEX OR J PLANT ECOL OR J CHEM ECOL OR ECOL MODEL OR BASIC APPL ECOL OR BIODIVERS CONSERV OR ECOSPHERE OR EVOL ECOL OR AQUAT MICROB ECOL)