Received: 3 February 2016 | Revised: 2 October 2017 | Accepted: 9 October 2017

DOI: 10.1111/geb.12684

MACROECOLOGICAL METHOD

WILEY Global Ecology and Biogeography A Journal of Macroecology

Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics

Yoan Fourcade[1,2] | Aurélien G. Besnard[1,3] | Jean Secondi[1,4,5]

[1]GECCO (Group of ecology and conservation of vertebrates), Université d'Angers, Angers, France

[2]Department of Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden

[3]LPO Aquitaine, Villenave d'Ornon, France

[4]UMR CNRS 5023 LEHNA, University Lyon 1, Lyon, France

[5]UMR CNRS 6554 LETG-LEESA, Université d'Angers, Angers, France

Correspondence
Yoan Fourcade, Department of Ecology, Swedish University of Agricultural Sciences, Box 7044, 75007, Uppsala, Sweden.
Email: yoanfourcade@gmail.com

Funding information
Plan Loire Grandeur Nature; European Regional Development Fund (ERDF); Région des Pays de la Loire; Agence de l'eau Loire-Bretagne; Angers Loire Métropole; Direction Régionale de l'Environnement; de l'Aménagement et du Logement (DREAL) du bassin de la Loire; Département du Maine-et-Loire

Editor: Michael Borregaard

Abstract

Aim: Species distribution modelling, a family of statistical methods that predicts species distributions from a set of occurrences and environmental predictors, is now routinely applied in many macroecological studies. However, the reliability of evaluation metrics usually employed to validate these models remains questioned. Moreover, the emergence of online databases of environmental variables with global coverage, especially climatic, has favoured the use of the same set of standard predictors. Unfortunately, the selection of variables is too rarely based on a careful examination of the species' ecology. In this context, our aim was to highlight the importance of selecting ad hoc variables in species distribution models, and to assess the ability of classical evaluation statistics to identify models with no biological realism.

Innovation: First, we reviewed the current practices in the field of species distribution modelling in terms of variable selection and model evaluation. Then, we computed distribution models of 509 European species using pseudo-predictors derived from paintings or using a real set of climatic and topographic predictors. We calculated model performance based on the area under the receiver operating curve (AUC) and true skill statistics (TSS), partitioning occurrences into training and test data with different levels of spatial independence. Most models computed from pseudo-predictors were classified as good and sometimes were even better evaluated than models computed using real environmental variables. However, on average they were better discriminated when the partitioning of occurrences allowed testing for model transferability.

Main conclusions: These findings confirm the crucial importance of variable selection and the inability of current evaluation metrics to assess the biological significance of distribution models. We recommend that researchers carefully select variables according to the species' ecology and evaluate models only according to their capacity to be transfered in distant areas. Nevertheless, statistics of model evaluations must still be interpreted with great caution.

KEYWORDS
AUC, environmental predictors, environmental variables, MaxEnt, model evaluation, ROC curve, species distribution modelling, TSS

1 | INTRODUCTION

Species distribution models (SDMs) have become in recent years one of the most widely used tools in macroecology. The principle of SDMs is to correlate species occurrences and environmental layers to build statistical inferences about the processes driving species' niches, and eventually derive suitability maps (Elith & Leathwick, 2009). This family of methods has a broad range of applications, including the study of niche evolution (Warren, Glor, & Turelli, 2008) and the delineation of conservation areas (Esselman & Allan, 2011). When models are projected in space and time, they can predict range shifts under climate change (Hijmans & Graham, 2006) or estimate the potential expansion of invasive species (Jiménez-Valverde, Peterson et al., 2011). The availability of large biodiversity (Edwards, Lane, & Nielsen, 2000) and environmental (Hijmans, Cameron, Parra, Jones, & Jarvis, 2005) databases, in combination with the release of user-friendly and powerful SDM

algorithms, has stimulated an exponential growth of SDM studies (Lobo, Jiménez-Valverde, & Hortal, 2010). As it is now a standard tool to inform conservation decisions (Guisan et al., 2013), it is critical to ensure that key aspects of the modelling workflow, such as variable selection and model performance assessment, can be adequately controlled and evaluated.

The choice of environmental predictors to include in the model is a crucial step in setting up SDMs. Machine-learning algorithms can in theory handle a large number of predictors and identify which ones are important via regularization (Elith et al., 2011). However, the output of SDMs can vary substantially depending on whether a subset of variables is pre-selected or not, especially when models are projected into new environments (Rödder, Schmidtlein, Veith, & Lötters, 2009; Synes & Osborne, 2011). A prevailing recommendation is to explore the correlation between predictors and select them to avoid multicollinearity (Braunisch et al., 2013). Generally, SDMs built from variables that are only incidentally or indirectly linked to a species' distribution can successfully fit their present-day range (Dormann et al., 2012). However, if the objective is to reveal the environmental drivers underlying a species' distribution, or to transfer the model to new areas, rigorously selecting relevant predictors according to the species' ecology is imperative (Petitpierre, Broennimann, Kueffer, Daehler, & Guisan, 2017).

It is generally assumed that climate is the main driver of species distributions at large spatial scales (Soberón, 2007). The 19 bioclimatic variables available as part of the Worldclim project (Hijmans et al., 2005) provide a source of supposedly biologically relevant climate data that can be integrated readily into SDM workflows. Booth, Nix, Busby, and Hutchinson (2014) found that, among the studies that implemented MaxEnt (Phillips, Anderson, & Schapire, 2006) as the SDM algorithm, 76% used bioclimatic variables as environmental predictors and 55% used all the 19 variables. The same pattern was observed by Bradie and Leung (2017) and Porfirio et al. (2014) who identified bioclimatic variables and elevation as the most commonly employed predictors in SDM studies. Although this standard set of 19 bioclimatic variables is frequently selected for modelling species' distributions, it is recognized that climatic factors are unable to describe in all their complexity the processes that limit species' ranges (Pearson & Dawson, 2003). It has also been suggested that the apparent correlation between climate and species' distributions may partly reflect the spatial structure of climate rather than a real biological process (Bahn & Mcgill, 2007; Beale, Lennon, & Gimona, 2008; Chapman, 2010). As it can be difficult to anticipate precisely the factors that drive a species' distribution, it is essential that the performance and the biological significance of SDMs can be evaluated to avoid drawing inferences from irrelevant environmental variables.

Post-hoc evaluation of distribution models is commonly performed to assess their predictive performance and statistical significance (Peterson et al., 2011). The most common diagnostic metrics in the area of SDMs is the area under the receiver operating curve (ROC) (AUC; Porfirio et al., 2014), obtained by plotting the model sensitivity against its false positive rate at all possible thresholds (Hanley & McNeil, 1982). Originally developed to assess the discrimination ability of radar systems or medical diagnostics, it has been widely adopted by the SDM community to measure the performance of models in discriminating between presences and absences of species (Lobo, Jiménez-Valverde, & Real, 2008). AUC has been adapted to presence-only (or presence–background) modelling approaches by comparing the predicted suitability at presence points versus background points taken from the training area. In this context, the implementation of AUC in a SDM framework is usually carried out by partitioning species occurrences into two sets: a training dataset, which is used to compute the model, and a test dataset that is used thereafter to evaluate the model's discrimination ability (Fielding & Bell, 1997). This process can be repeated several times, each partition being used alternately to train and to test the model. This approach assumes that training and testing data are spatially independent, an assumption rarely fully met in practice, especially when occurrences are randomly partitioned (Veloz, 2009). Moreover, AUC has been recognized as a highly questionable measure for several years (Lobo et al., 2008), especially when used with background data instead of true absences (Jiménez-Valverde, 2012). Many alternative metrics have been proposed to evaluate SDMs (see for example Allouche, Tsoar, & Kadmon, 2006; Hijmans, 2012; Phillips & Elith, 2010). However, despite these criticisms, so far none of these alternatives seems to have taken over from AUC in most SDM studies.

In this context, the use of improper environmental predictors in SDMs may remain overlooked if the statistics used to assess their performance are unable to distinguish models with no biological realism. Here, our aim was to study how much these issues can interact to bias inferences based on species distribution models, and to investigate potential solutions to overcome them. We first reviewed the current practice in the field of SDMs in terms of variable selection and model evaluation. Second, we evaluated the performance of SDMs built from meaningless predictors according to classical SDM evaluation statistics [AUC and also the true skill statistics (TSS, Allouche et al., 2006), often recommended as an alternative to AUC]. Instead of computing spatial null models (Bahn & Mcgill, 2007; Beale et al., 2008) or simulated climate data (Chapman, 2010), our approach consisted of using completely meaningless variables derived from paintings, not selected based on any prior criterion. We then compared the evaluation of SDMs built using these pseudo-predictors with SDMs built using real environmental variables often used in empirical SDM studies (bioclimatic variables and elevation). When evaluating SDM outputs, we tested various approaches to partitioning occurrences that provide different levels of spatial independence between both, and assessed whether one of those strategies better discriminated meaningless models from those computed using real environmental predictors.

2 | METHODS

2.1 | Review of current practices in species distribution modelling

In order to characterize the current practices regarding variable selection and model evaluation in SDMs, we conducted a literature review of modelling papers published in the last few years in the four leading journals of macroecology and biogeography: *Global Ecology and*

Biogeography, *Journal of Biogeography*, *Diversity and Distributions* and *Ecography* (Figure 1). Specifically, we restrained our analysis to articles that cited at least one of the two main references for MAXENT, the most widely used SDM method: Phillips et al. (2006) and Phillips and Dudík (2008). We used Web of Science (® Thomson Reuters) to download all corresponding articles published from 2012 (queried on 22 June 2016) in the selected journals. We surveyed 246 articles, from which we manually excluded 36 because they were not empirical modelling studies or because they focused on evaluating the performance of various types of environmental predictors. All the 210 remaining studies were kept for the description of SDM evaluation methods but, as our aim was mainly to describe the use of bioclimatic variables, 20 additional studies on marine organisms were discarded for the analysis of variable choice. For the selected papers, we recorded the methods used to evaluate the performance of SDMs, the variables included in the models and how they were selected. Specifically, we noted whether authors selected their variables based on a statistical method (most often by selecting the least correlated ones) and whether they provided a clear biological justification of the choice of the environmental predictors. For the latter criterion, we considered predictor selection as being biologically motivated if it was associated with a description of the species' biology/ecology, and supported by references. Studies that provided only a vague statement such as 'we believe that these variables are relevant for the species', or 'climate is thought to be the main driver of species' distributions' were classified as being partially justified. We additionally performed a correspondence analysis to describe the differences between journals.

2.2 | Species occurrences datasets

We used the datasets from 497 species listed on the European Red List, an assessment of the conservation status of *c.* 6000 Western Palaearctic species. Occurrences were downloaded from the Global Biodiversity Information Facility database (GBIF, http://www.gbif.org). We selected species for which between 500 and 2000 presence points were available, and kept only the records with valid coordinates. To avoid sampling bias that may affect the models' outputs (Syfert, Smith, & Coomes, 2013), we used the procedure implemented in the spThin (Aiello-Lammens, Boria, Radosavljevic, Vilela, & Anderson, 2015) R package, which consists of a spatial thinning of occurrence records, or spatial filtering (Fourcade, Engler, Rödder, & Secondi, 2014). Using a randomization approach, spThin returned a dataset containing the maximum number of occurrences separated by a user-defined minimum neighbour distance. The optimal distance that substantially reduces sampling bias while keeping a sufficient amount of information is challenging to assess, especially as we dealt with a large number of species. Thus, we used a conservative measure for all species, defined as the maximum distance between two occurrences divided by 20. This filtering distance appeared visually to provide a reasonable trade-off at various spatial scales (Supporting Information Figure S1). The final set of 497 species and their number of occurrences before and after spatial filtering (mean = 44; min = 10; max = 245) are given in Supporting Information Table S1.

2.3 | Environmental variables and pseudo-predictors

We first aimed at building a model based on predictors that were potentially biologically relevant and widely used in SDM studies (Booth et al., 2014; Bradie & Leung, 2017; Porfirio et al., 2014 and see Results). We used the full set of 19 bioclimatic variables and the altitudinal grid available from the Worldclim database (Hijmans et al., 2005) at 10 arc-minute resolution and rescaled to 20 arc-minute resolution (Supporting Information Figure S2). As we aimed at modelling distributions at the European scale, each grid was cropped between −10 to 70° longitude and 30 to 75° latitude.

A set of 20 pseudo-predictors (Figure 2 and Supporting Information Figure S3) was additionally created. We downloaded image files by performing a request on the Google Image® search engine (http://images.google.com) with the term 'classical paintings', searching only files with a 'large size' according to the search tool. Twenty files, in jpeg format, were selected among the first hits. Each image file was mapped on the European geographical space using the following protocol: files were imported into R 3.0.2 (R Development Core Team, 2015) using the 'raster' function in the eponymous package (Hijmans, 2014). By default, this function imports the red component of images, coded as values ranging from 0 to 255. Thereafter, we matched the spatial extent and resolution of each picture with the real environmental predictors using the 'extent' and 'resample' functions of the raster R package. Finally, a mask was applied to crop each of these pseudo-predictors to European land surfaces.

2.4 | Species distribution modelling

In order to avoid the problems associated with multicollinearity among predictors (Braunisch et al., 2013), and as we are not interested in the variables' responses, we computed a principal components analysis (PCA) for each dataset (Dormann et al., 2013). As the 12 first principal components were needed to explain at least 80% of the paintings-based predictors' variance, we ran SDMs for each species using as environmental predictors the 12 first PCA axes derived from: (a) the 20 real environmental predictors and (b) the 20 pseudo-predictors derived from the paintings. In addition, to test how much the number of bioclimatic variables included in the models affected our results, we also computed SDMs for 100 randomly selected species using 2 to 19 predictors (keeping the least correlated ones), both for bioclimatic variables and for pseudo-predictors.

We used the method implemented in MAXENT version 3.3.3k (Phillips et al., 2006), with species-specific settings selected using the ENMeval (Muscarella et al., 2014) R package. The approach implemented in ENMeval runs successively several MAXENT models using different combinations of parameters to select the settings that optimize the trade-off between goodness-of-fit and overfitting. Here, we set ENMeval to test regularization values between 0.5 and 4, with 0.5 steps, as well as the following feature classes: linear, linear + quadratic, hinge, linear + quadratic + hinge, linear + quadratic + hinge + product and linear + quadratic + hinge + product + threshold, which corresponds to the default ENMeval settings. The extent of the training area, that is, the portion of environmental grids from which background points are

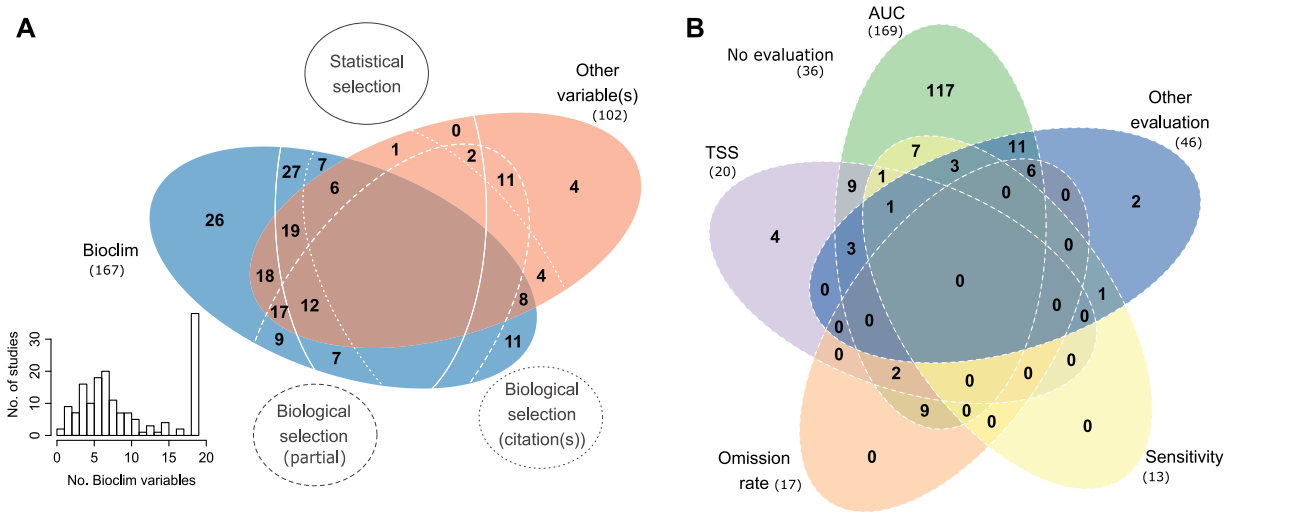

FIGURE 1 Venn diagrams illustrating current practices in species distribution modelling (data from articles that cite MAXENT in the four leading journals of macroecology and biogeography, from 2012). (a) Use of bioclimatic variables (blue ellipse) versus all other types of variables (orange ellipses), and variable selection. Dotted and dashed ellipses: biologically motivated selection; plain white ellipse: statistical selection. The distribution of the number of bioclimatic variables incorporated in studies that used them is shown in the bottom-left inset. (b) Use of various evaluation metrics, including area under the receiver operating curve (AUC; green ellipse) and metrics that were present in more than 5% of the reviewed papers (other methods are grouped under the blue ellipse). Note that due to the number of groups and the presence of zeros, the area of each ellipse is not proportional to the number of cases. TSS = true skill statistics

sampled, should correspond to the extent actually accessible for the species, so that its absence in this region reflects the effect of environmental factors (Barve et al., 2011). In order to keep a comparable definition of the training area among species, we created buffers around all occurrences with a radius equal to half the maximum distance between points. Two thousand points were selected from within the area of each buffer and used as background points in the ENMeval workflow. All MAXENT models were run with the 48 combinations of settings, and we selected the one that had the lowest corrected Akaike information criterion (AICc; Burnham & Anderson, 2002) for further analyses.

2.5 | Training/testing partitioning and model evaluation

SDM evaluation was performed by partitioning occurrences into training and testing datasets in order to report the AUC both for models computed with real environmental variables (AUCe) and with paintings-derived pseudo-predictors (AUCp). AUC is 0 when there is a total mismatch between model predictions of presence and the actual data and 1 for models with perfect discrimination abilities. A value of 0.5 indicates that the model does not perform better than any model with a set of random predictors. Additionally, we computed the TSS (also known as the Youden index), a threshold-dependent evaluation metric (Allouche et al., 2006; Youden, 1950), calculated as sensitivity + specificity − 1. The threshold for converting continuous maps to binary predictions was set to maximize the model specificity and sensitivity, as recommended by Liu, White, and Newell (2013). Again, values were reported for models computed with environmental (TSSe) and paintings (TSSp) variables. In each case, the occurrence dataset was divided into four bins, used in a cross-validation approach where each bin was used in turn as test points while the three others were used to train the model. Evaluation metrics were then averaged across the four possible pairs of training/test datasets.

In order to test the effect of various methods of training/testing partitioning, we adopted four approaches that provided different levels of spatial independence between training and testing datasets (Figure 3). First, we simply randomly divided occurrences into four groups. Second, we performed three types of spatially structured partitioning. The two first ones are variations of the 'checkerboard' approach implemented in ENMeval (Muscarella et al., 2014): the geographical space was first transformed into a checkerboard-like grid by applying an aggregation factor to the original grid resolution. A second, twofold coarser, independent checkerboard-like grid, was then applied on top of the first one. The combination of both levels of partitioning allowed the division of occurrences into four bins [see Muscarella et al. (2014) and Figure 3 for more details]. This 'checkerboard' approach was tested for two spatial grains: with an aggregation factor of two as ENMeval uses this as the default value (fine-resolution checkerboard), and with an aggregation factor of four (coarse-resolution checkerboard). Finally, we implemented the 'block' approach of ENMeval that partitions occurrences according to their longitude and latitude, as recommended by Radosavljevic and Anderson (2014). It results in four geographically non-overlapping bins of equal numbers of occurrences, corresponding to each corner of the geographical space. In this approach, as well as in both 'checkerboard' approaches, background points were also split following the same spatial partitions (corners or checkerboard cells). Then, at each modelling step, the model was trained without thebackground points located in the same area as the test points. The 'block' method provides the best spatial independence between training and testing datasets that can be obtained from partitioning a unique dataset. Therefore, this method quantifies the ability of the models to extrapolate their predictions into new areas.

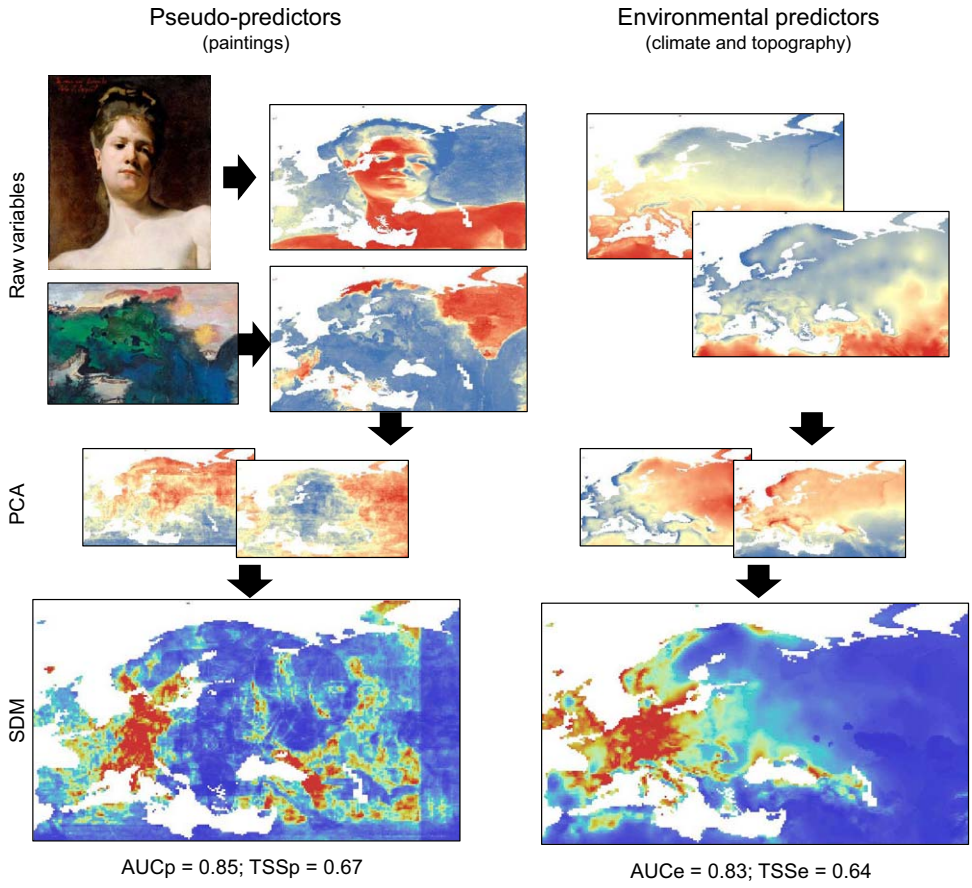

FIGURE 2 Workflow used in analyses: 20 pseudo-predictors were created from the projection of paintings on the Western Palaearctic geographical space (examples: top: John Singer Sargent, *Blonde Model*, bottom: Zhang Daqian, *Spring dawns upon the colorful hills*) and were used to compute species distribution models (SDMs) after principal components analysis (PCA). A set of 20 true environmental variables (climate and topography) was also used to compute SDMs for the same species. Both types of models were evaluated using area under the receiver operating curve (AUC) and true skill statistics (TSS). The SDMs presented at the bottom show the example of a species (*Candidula unifasciata*, a land snail species) for which the SDM computed with pseudo-predictors led to better evaluation metrics (here computed by randomly splitting occurrences into training and testing datasets) than that computed with real environmental variables (suitability increases from blue to red). AUCp = AUC for model computed with paintings-derived pseudo-predictors; AUCe = AUC for model computed with real environmental variables; TSSp = TSS for model computed with paintings-derived pseudo-predictors; TSSe = TSS for model computed with real environmental variables

We used repeated measures ANOVA and post-hoc Tukey tests to test whether AUC and TSS values, and the differences between model evaluations based on paintings and real environmental variables, significantly differed between these different partitioning strategies. We also reported the correlation between AUCe and AUCp, and between TSSe and TSSp, using linear regressions.

3 | RESULTS

3.1 | Current practices in species distribution modelling

Among the 190 studies that modelled terrestrial organisms, 167 (87.9%) included at least one of the 19 classical bioclimatic variables, including 38 (20%) that used all the 19 variables. A total of 102 articles (53.7%) used another set of variables instead of (23, 12.1%) or in addition to (80, 42.1%) bioclimatic variables (Figure 1a). These other variables included, for example, predictors linked to topography (the most common: 55 studies, 28.9%), other climatic variables (23, 12.1%), vegetation cover or productivity (21, 11.1%), land cover (21, 11.1%) or soil characteristics (20, 10.5%). The selection of the environmental variables was biologically motivated in 105 papers (50.0%), including 64 (30.5%) that provided at least one citation to justify the inclusion of one or several variables, and 41 other articles (19.5%) that provided only a vague justification. Moreover, 88 studies (41.9%) selected variables based on a statistical criterion (correlation between predictors or occasionally model selection) instead of, or in addition to, selection based on the biology/ecology of the species (Figure 1a).

Among all the 210 scanned articles, 169 (80.5%) used the AUC to evaluate SDMs, including 117 (55.7%) that used only AUC (Figure 1b). The next most commonly used evaluations were three threshold-dependent metrics: the TSS (20, 9.5%), the omission rate (17, 8.1%) and the model sensitivity (13, 6.2%). Various other methods of evaluation, including for example Kappa, specificity, commission rate or overall

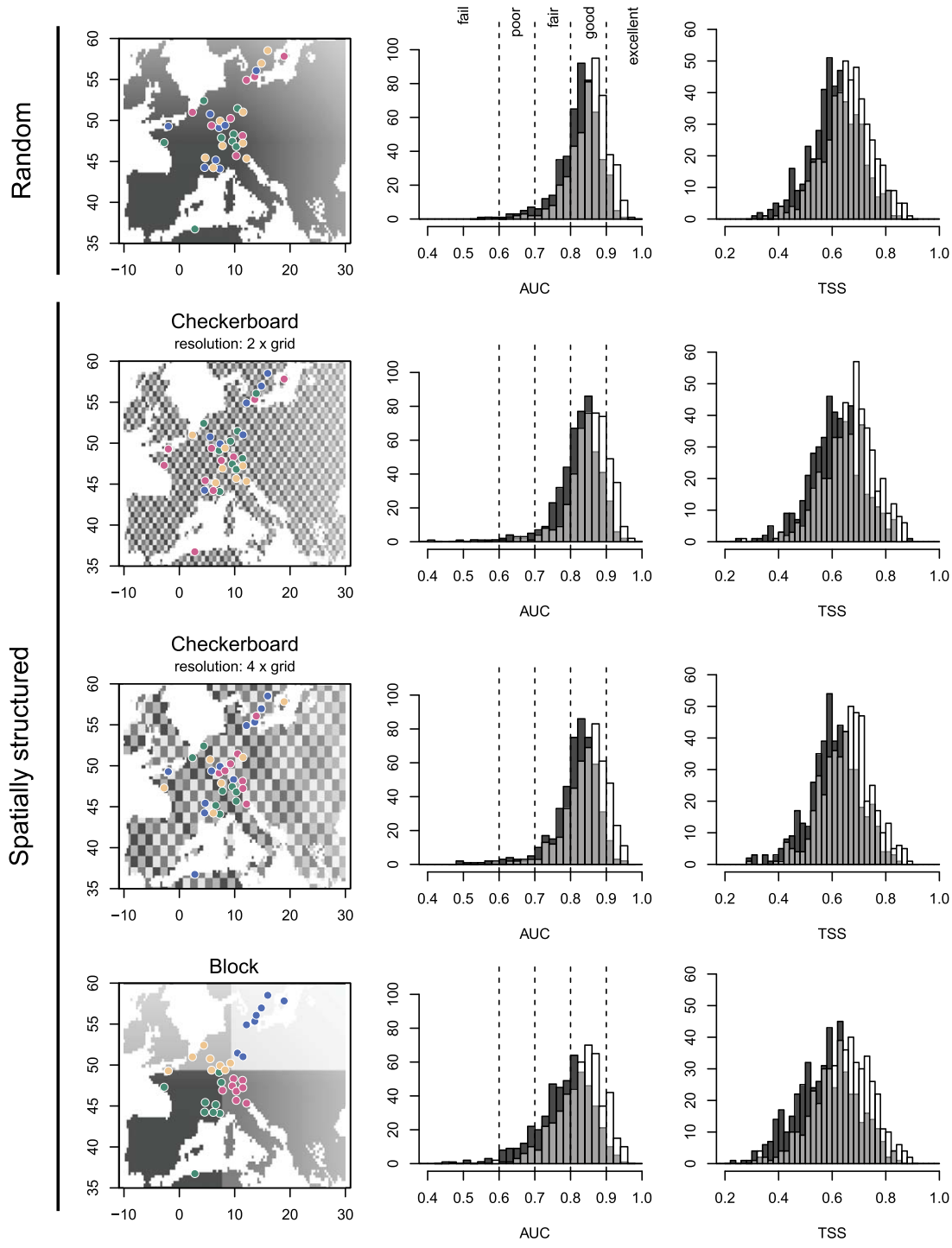

FIGURE 3 Schematic representation of the four approaches for partitioning occurrences into training and testing datasets (left column). For spatially structured methods, the greyscale grid shows the division of the geographical space that defines how occurrences are split into four bins. The central and right columns show the corresponding distribution of evaluation metrics [area under the receiver operating curve (AUC) and true skill statistics (TSS), respectively] for models trained using pseudo-predictors derived from paintings (dark grey) or from real environmental variables (white, with overlap in light grey). Vertical dotted lines on top of AUC distributions represent the usual classification of AUC by Araújo et al. (2005), adapted from Swets (1988)

accuracy, were present in less than 5% of all articles reviewed. About a quarter of all articles (53, 25.2%) used several evaluation measures, 58.6% (123) used only one method of evaluation and 16.1% (34) did not evaluate their models, or at least did not report it (Figure 1b).

There was no marked difference in terms of variable selection and model evaluation between journals (Supporting Information Figure S4). However, a noticeable pattern is the tendency for articles published in *Journal of Biogeography* to use on average a higher number of

bioclimatic variables compared to the other journals. For example, among the 25 studies (13.2%) that used the whole set of 19 bioclimatic variables as only predictors, 20 were published in *Journal of Biogeography*. In addition, all but one of the 15 studies that used the 19 bioclimatic variables as the only predictors and AUC as the only evaluation measure were published in *Journal of Biogeography*.

3.2 | Effect of the type of predictor and the partitioning approach on model evaluation

Overall, the performance of SDMs computed from real environmental variables differed depending on the approach used to partition occurrences into testing and training datasets, according to both AUCe ($F_{3,1488} = 69.09$, $p < .001$) and TSSe: ($F_{3,1488} = 21.13$, $p < .001$). Post-hoc comparisons (Tukey honest significant difference (HSD) tests) showed that the random (mean AUCe = 0.85 ± 0.0026 *SEM*, mean TSSe = 0.66 ± 0.0043 *SEM*) and fine-resolution 'checkerboard' (mean AUCe = 0.85 ± 0.0029 *SEM*, mean TSSe = 0.66 ± 0.0045 *SEM*) methods led to similar AUCe ($p = 1$) and TSSe ($p = 1$). Using the standard classification of AUC of Araújo, Pearson, Thuiller, and Erhard (2005), adapted from Swets (1988), more than 85% of these models were classified as 'good' (AUCe > 0.8) or 'excellent' (AUCe > 0.9) (Figure 3). However, models evaluated using the 'block' method had significantly lower AUCe (mean = 0.83 ± 0.0032 *SEM*) and TSSe (mean = 0.64 ± 0.0048 *SEM*) than all other methods ($p < .001$ for all pairwise comparisons), and only 68% of them were classified as 'good' or 'excellent' based on their AUCe. Evaluations provided by the coarse-resolution 'checkerboard' approach were intermediate. AUCe (mean = 0.84 ± 0.0028 *SEM*) was in this case significantly higher than for the 'block' method ($p < .001$) and lower than for the fine-resolution 'checkerboard' and the random partitioning methods ($p = 0.01$ and $p = 0.001$, respectively), which led to 81% of models being classified at least as 'good'. TSSe (mean = 0.65 ± 0.0043 *SEM*), however, was significantly higher than for models evaluated with the 'block' method ($p < .001$), but did not differ with other partitioning approaches ($p = .06$ with the fine-resolution 'checkerboard' method and $p < .001$ with the random partitioning).

Models trained from pseudo-predictors showed the same pattern: AUCp and TSSp values differed depending on the partitioning method (AUCp: $F_{3,1488} = 138.87$, $p < .001$, TSSp: $F_{3,1488} = 58.73$, $p < .001$). It was again driven by a lower evaluation score of SDMs with the 'block' method (mean AUCp = 0.78 ± 0.0036 *SEM*, mean TSSp = 0.57 ± 0.0055 *SEM*, $p < .001$ compared to all other methods) that led to 48% of models being classified at least as 'good' (Figure 3). In contrast, the random (mean AUCe = 0.82 ± 0.0029 *SEM*, mean TSSp = 0.61 ± 0.0047 *SEM*) and fine-scale 'checkerboard' (mean AUCe = 0.82 ± 0.0031 *SEM*, mean TSSe = 0.61 ± 0.0049 *SEM*) methods did not differ in their AUCp ($p = 1$) or TSSp ($p = 1$), and 71 and 74%, respectively, of these models were classified as 'good' or 'excellent'. Again, the coarse-scale 'checkerboard' evaluation led to intermediate AUCp (mean = 0.81 ± 0.0034 *SEM*; $P < 0.001$ with all other methods; 69% of models had AUCp > 0.8) and TSSp values (mean = 0.60 ± 0.0047 *SEM*; $p = .07$ with the fine-resolution 'checkerboard' method, $p = .28$ with the random partitioning and $p = .06$ with the 'block' method).

AUCe and AUCp, as well as TSSp and AUCp, were significantly correlated in all cases (Supporting Information Figure S5), but this relationship was weaker when models were evaluated with the 'block' approach (AUC: $R^2 = .06$; TSS: $R^2 = .11$, compared to $R^2 > .15$ for other partitioning methods). Even though on average, AUCp and TSSp values were lower than AUCe and TSSe (Wilcoxon tests, $p < .001$ for all approaches), whatever the partitioning method, around 30% of models had a higher AUC and TSS when pseudo-predictors rather than real environmental variables were used. However, for a given species, AUCe was higher than AUCp by 6.5% (± 0.63% *SEM*) on average using the 'block' method of occurrence partitioning (Figure 4). This was significantly higher than the average of 4.2% (± 0.44% *SEM*), 4.4% (± 0.47% *SEM*) and 4.9% (± 0.53% *SEM*) for the random and the fine- and coarse-resolution 'checkerboard' approaches, respectively (ANOVA: $F_{3,1488} = 14.98$, $p < .001$, post-hoc Tukey tests: p always $< .001$). Pairwise comparisons showed however that the difference between AUCp and AUCe was similar for the random and both 'checkerboard' approaches ($p > .36$; Figure 4). On the contrary, the difference in TSS between models based on real environmental variables and from pseudo-predictors did not significantly differ between partitioning approaches (ANOVA: $F_{3,\ 1488} = 1.61$, $p = .19$). Still, the 'block' method tended to provide a higher (non-significant) improvement in TSSe compared to TSSp (27.8 ± 10.25% *SEM* vs. 13.7 ± 2.82% *SEM*, 15.5 ± 3.64% *SEM* and 16.3 ± 4.08% *SEM* for the random and the fine- and coarse-resolution 'checkerboard' approaches, respectively; Figure 4).

The number of predictors included in the models did not dramatically affect the results, although the patterns described above were stronger when the number of predictors increased (Supporting Information Figure S6). The measures of SDM performance tended to increase with the number of predictors, whatever the partitioning method or the type of predictors. On average, AUCe and TSSe were consistently higher than AUCp and TSSp but the more predictors that were included, the more this difference decreased (excluding models with only two variables whose evaluations were similarly low for both types of predictors). Similarly, at least 11% of models were better evaluated when computed using pseudo-predictors, but this proportion increased with the number of predictors [around 25% (block partitioning) or 30% (random partitioning) for models with more than 15 predictors].

4 | DISCUSSION

We have demonstrated that SDMs computed using meaningless variables as input environmental predictors are often classified as good or even excellent according to the most widely used evaluation measures. Worse, one third of these models led to better evaluations than those computed from the real environmental predictors that most studies utilize. This result shows that SDMs with no biological realism could easily occur and remain undetected, and questions the current practices of the SDM community in terms of model evaluation and variable selection.

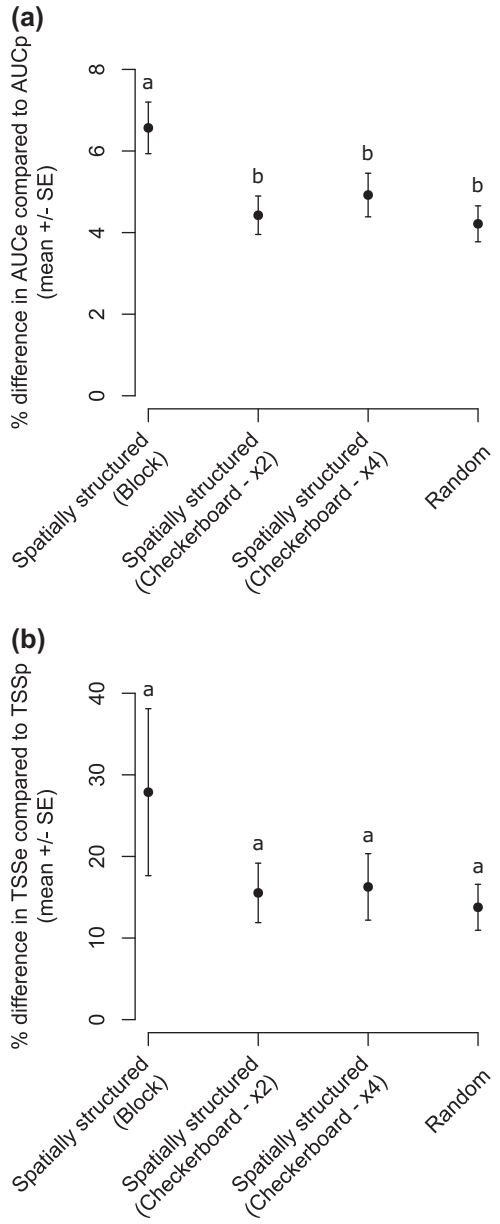

FIGURE 4 Mean difference (in %, ± *SE*) in evaluation metrics [area under the receiver operating curve (AUC): (a); true skill statistics (TSS): (b)] between models built from true environmental variables or pseudo-predictors derived from paintings. Significant differences between groups (Tukey's post-hoc tests after repeated-measures ANOVA) are shown by letters (see text for details). AUCp = AUC for model computed with paintings-derived pseudo-predictors; AUCe = AUC for model computed with real environmental variables; TSSp = TSS for model computed with paintings-derived pseudo-predictors; TSSe = TSS for model computed with real environmental variables

Although on average AUC and TSS tended to be higher for climate-based models than for models built from pseudo-predictors, in a large proportion of cases the latter were at least equally well evaluated. This implies that, either bioclimatic and topographic variables are no better at predicting species distributions than any randomly chosen set of non-biological variables, or classical evaluation metrics are unable to identify wrong models. The first hypothesis draws upon recent work by Bahn and Mcgill (2007), Beale et al. (2008) and Chapman (2010) who suggested that the apparent association between climate and species' ranges may be equally well explained by the inherent spatial autocorrelation of species distributions as by an actual biological process. In contrast to these studies, we did not attempt to create a spatial null model to be compared to climatic models. Instead, we used randomly selected paintings as pseudo-predictors, but in essence, the approach is similar. Any drawing or environmental raster exhibits autocorrelated spatial patterns that can be statistically fitted to a dataset of species' occurrences, especially when the number of predictors is high. Therefore, a parallel spatial autocorrelation between a species' presence and environmental variables can make it hard to distinguish between geographical patterns and biological processes (Warren, Cardillo, Rosauer, & Bolnick, 2014). We note, however, that many historical (von Humboldt & Bonpland, 1805; Wallace, 1876) and modern (Araújo & Peterson, 2012; Jiménez-Valverde, Barve et al., 2011; Thomas, 2010) empirical pieces of evidence demonstrate that climate plays a primary role in shaping species' distributions not least of which is the range shift observed in many species in synchrony with climate change (Chen, Hill, Ohlemuller, Roy, & Thomas, 2011).

If we make the reasonable assumption that paintings are not as good predictors of species distributions as bioclimatic and topographic variables, then we have to question the reliability of the metrics we used to evaluate the predictive performance of models. Previous studies have already shown that AUC cannot provide an efficient measure of SDM performance (Jiménez-Valverde, 2014; Lobo et al., 2008; Smith, 2013), and especially that it does not inform about models' biological significance (Fourcade et al., 2014; Stolar & Nielsen, 2015). The drawbacks of AUC, as a measure to assess SDMs, have been attributed to its dependence on the calibration area (Barve et al., 2011; Jiménez-Valverde, Acevedo, Barbosa, Lobo, & Real, 2013) as well as to the fact that it ignores the spatial distribution of errors and that it relies on the ranking of sensitivity and specificity across thresholds, ignoring the actual probability values given by the model (Lobo et al., 2008). Moreover, AUC tends to inflate for models that have a strong fit to input presence points, and thus favours those that estimate realized distribution while penalizing those that predict the species' potential distribution (Jiménez-Valverde, 2012). It also weights equally omission and commission errors, a property that is not necessarily desirable (Jiménez-Valverde, 2012, 2014; Lobo et al., 2008). Threshold-dependent statistics like sensitivity and specificity have been suggested as valuable alternatives to evaluate the discrimination ability of models (Jiménez-Valverde, 2014). In this regard, our results revealed that TSS, which is basically a recapitulation of sensitivity and specificity for a given threshold, is similarly unable to assess the predictive value of the input variables. Here, we were unable to calculate the true AUC and specificity as they were based on background data instead of true absences. Possibly, evaluation metrics based on real presence/absence data may perform better (Jiménez-Valverde, 2012; Smith, 2013). However, presence/background modelling has become standard as most SDMs are trained from occurrence records only. Providing that true absences and independent test data are available, interpreting the entire ROC curve – or its most relevant section (Peterson, Papeş, &

Soberón, 2008) – instead of summarizing it by a single value may give more useful insights into SDM discrimination ability (Jiménez-Valverde, 2012). However, such an ideal set-up is most often out of reach in real-world situations, and in any case a careful examination of ROC plots may be too complex for studies involving multiple species. Even if evaluation statistics can inform about model fit, it is unlikely that a single value can tell whether a SDM has any biological realism or not. (Araújo et al., 2005).

The flaws of AUC that our results suggest have been known for a while (Lobo et al., 2008). Yet, it is still used in more than 80% of distribution modelling papers published in recent years in leading biogeography journals. More than half of SDM studies even relied on this single measure alone to assess the performance of their models. At the same time, the large majority of these studies included all or some of the 19 bioclimatic variables popularized by the Worldclim project (Hijmans et al., 2005). Based on the assumption that climate must play a role in driving species distributions, they have become a standard default predictor set in most modelling studies. Some disciplines such as palaeobiogeography or phylogeography have no choice but to rely on climate models, as they are the only environmental predictors of past distributions available across geological scales. However, the importance of climate in shaping species distributions is often unevaluated and a clear justification of this choice of variables is frequently lacking. If evaluation measures fail to identify when distribution models are built from obviously meaningless predictors, how much can we trust the many studies that make use of SDM approaches? The answer probably depends on the objective of the study and on the weight given to the evaluation of model performance. Although interpretations of model performance based on AUC or even TSS are most likely misleading, it is probably safe to adopt a relatively relaxed selection of bioclimatic predictors when the aim is explicitly to model climatically suitable areas and not necessarily the realized distribution of a species. Similarly, environmental variables that are only proxies or that are spatially correlated to direct predictors of species presence may be sufficient to approximate a species' current distribution. However, problems arise when models are projected in space or time. Indeed, real descriptors of the causal factors that determine the distribution of the species are needed when SDMs are used to hindcast palaeodistributions (Varela, Lobo, & Hortal, 2011) or to forecast range shifts with climate change (Austin & Van Niel, 2011). The same requirement applies to models aimed at predicting the invasive potential of species in a different area than the calibration range (Petitpierre et al., 2017).

As it appears that standard evaluation metrics can hardly discriminate biologically relevant SDMs from meaningless models, the question now is whether strategies exist that can overcome this problem. First of all, the importance of carefully selecting predictors according to the known ecology or physiology of the species of interest must be emphasized (Austin & Van Niel, 2011; Petitpierre et al., 2017; Rödder & Lötters, 2010; Rödder et al., 2009). Involving expert knowledge in the process of variable selection can help to identify a priori the true drivers of species distribution (Murray et al., 2009). As the risks of using strongly correlated variables have also been recognized (Braunisch et al., 2013), many authors select variables according to their level of intercorrelation, a safe practice from a statistical point of view but which does not ensure the use of biologically interpretable predictors. There is also a growing interest in adapting information theoretic approaches inspired by procedures of model selection by AIC (Burnham & Anderson, 2002). The idea is to compare models computed with different sets of predictors and to select the best model according to criteria that account for model fit while penalizing overfitting (Warren & Seifert, 2011). Although it cannot replace a careful procedure of variable selection based on ecological criteria, it has been suggested that information criteria do not suffer from the drawbacks of AUC and might be a way to identify a set of relevant predictors for the modelled species. Methods of variable selection by AIC (and AICc or Bayesian Information Criterion (BIC)) have been implemented by several authors (Verbruggen et al., 2013; Warren, Glor, & Turelli, 2010; Zeng, Low, & Yeo, 2016) and have been proven to improve the performance of SDMs compared with models that use unselected variable sets. However, others had pointed out before that AIC is sensitive to spatial autocorrelation (Diniz-Filho, Rangel, & Bini, 2008). In this regard, our results showed that, as for AUC, *c*. 30% of models based on pseudo-predictors were better evaluated (lower AICc) than those based on real environment variables.

In addition to rigorous selection of variables, inferences based on SDMs could be improved by using an effective method of model evaluation. We showed, in line with other authors (Jiménez-Valverde, 2012; Lobo et al., 2008), that two widely used evaluation metrics overrate the performance of biologically meaningless SDMs. However, an essential assumption of these approaches that we did not discuss is the strict independence between training and test records (Araújo et al., 2005). In practice, occurrences of a species can rarely meet this assumption, especially when training and evaluation points are generated from the same dataset via cross-validation. A general strategy to increase the spatial independence of training and testing datasets is to split the data into spatial blocks (Roberts et al., 2017). This type of cross-validation implies that the spatial transferability of the model is evaluated rather than just its interpolation accuracy (Bahn & McGill, 2013; Wenger & Olden, 2012). It usually results in a lower evaluation of performance, but closer to an actual estimate of model transferability (Roberts et al., 2017). We found the same pattern in our study. The 'block' partitioning method, which consisted of dividing occurrences into four geographically separated bins (Radosavljevic & Anderson, 2014), led to lower evaluation measures for both sets of variables. Interestingly, AUCe and AUCp were more different and less correlated when they were estimated from this method compared to other partitioning strategies. We observed a similar but non-significant pattern for TSS as well. The coarse-resolution 'checkerboard' method, which also provided a certain level of spatial independence between training and test data, tended to show similar – although not significant – patterns. Evaluations using these methods remain remarkably high for SDMs trained from purely non-biological variables, and are thus potentially misleading on their own. They also failed to identify models based on irrelevant predictors as often as the others did. Still, these results suggest that highly unrealistic models tended to be penalized when their performance was evaluated via a spatial segregation of test and training occurrences. Our

results thus confirm previous findings that it is advisable to give priority to evaluation approaches that use spatially independent occurrences located in a distant area from the calibration dataset (Bahn & McGill, 2013; Radosavljevic & Anderson, 2014, Roberts et al., 2017; Wenger & Olden, 2012).

SDM workflows have too often consisted of downloading occurrence data from an online database and the 19 Worldclim bioclimatic variables (Hijmans et al., 2005), and building a model using the default settings in MaxEnt (Morales, Fernandez, & Baca-Gonzalez, 2017). However, owing to the many potential issues that may arise when settings and inputs are not adequately chosen (Merow, Smith, & Silander, 2013; Yackulic et al., 2013), such a simplistic approach can be considered as bad practice. Fortunately, SDM methods are not in their infancy anymore and awareness of these problems has risen. Nevertheless, a large proportion of studies still relies on inappropriate evaluation statistics and fails to report a rigorous selection of environmental predictors. We showed here that the inherent spatial autocorrelation of species distributions and environmental variables, associated with the use of non-independent test data and flawed evaluation metrics, could inflate the apparent performance of SDMs whatever their actual biological relevance. Therefore, in the absence of a robust method to evaluate the biological significance of SDMs, it appears essential to select a relevant set of environmental predictors, based on the known ecology or physiology of the species of interest. We also encourage SDM users to evaluate models in light of their transferability using spatially independent data, a crucial feature for most applications such as predicting the potential spread of invasive species (Rödder & Lötters, 2010; Verbruggen et al., 2013; Wang & Jackson, 2014) or forecasting climate-driven range shifts (Dobrowski et al., 2011; Elith & Leathwick, 2009). In any case, we recommend avoiding drawing strong conclusions about their performance based on metrics such as AUC or TSS. More research is still needed to assess the ability of different modelling algorithms to identify truly influential predictors, to develop reliable model evaluation measures, and to design statistical procedures of variable selection that could complement a biologically informed choice of predictors.

ACKNOWLEDGMENTS

The authors were financially supported by Plan Loire Grandeur Nature, European Regional Development Fund (ERDF), Région des Pays de la Loire, Agence de l'eau Loire-Bretagne, Angers Loire Métropole, Direction Régionale de l'Environnement, de l'Aménagement et du Logement (DREAL) du bassin de la Loire, and Département du Maine-et-Loire.

DATA ACCESSIBILITY

Model evaluation statistics for all species and partitioning approaches that form the basis of this manuscript are included in Supporting Information Table S1.

ORCID

Yoan Fourcade http://orcid.org/0000-0003-3820-946X
Jean Secondi http://orcid.org/0000-0001-8130-1195

REFERENCES

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., & Anderson, R. P. (2015). spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, *38*, 541–545.

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.

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

Araújo, M. B., Pearson, R., Thuiller, W., & Erhard, M. (2005). Validation of species–climate impact models under climate change. *Global Change Biology*, *11*, 1504–1513.

Austin, M. P., & Van Niel, K. P. (2011). Improving species distribution models for climate change studies: Variable selection and scale. *Journal of Biogeography*, *38*, 1–8.

Bahn, V., & Mcgill, B. J. (2007). Can niche-based distribution models outperform spatial interpolation? *Global Ecology and Biogeography*, *16*, 733–742.

Bahn, V., & McGill, B. J. (2013). Testing the predictive performance of distribution models. *Oikos*, *122*, 321–331.

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., … Villalobos, F. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, *222*, 1810–1819.

Beale, C. M., Lennon, J. J., & Gimona, A. (2008). Opening the climate envelope reveals no macroscale associations with climate in European birds. *Proceedings of the National Academy of Sciences USA*, *105*, 14908–14912.

Booth, T. H., Nix, H. A., Busby, J. R., & Hutchinson, M. F. (2014). BIOCLIM: The first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. *Diversity and Distributions*, *20*, 1–9.

Bradie, J., & Leung, B. (2017). A quantitative synthesis of the importance of variables used in MaxEnt species distribution models. *Journal of Biogeography*, *44*, 1344–1361.

Braunisch, V., Coppes, J., Arlettaz, R., Suchant, R., Schmid, H., & Bollmann, K. (2013). Selecting from correlated climate variables: A major source of uncertainty for predicting species distributions under climate change. *Ecography*, *36*, 971–983.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer-Verlag.

Chapman, D. S. (2010). Weak climatic associations among British plant distributions. *Global Ecology and Biogeography*, *19*, 831–841.

Chen, I. C., Hill, J. K., Ohlemuller, R., Roy, D. B., & Thomas, C. D. (2011). Rapid range shifts of species associated with high levels of climate warming. *Science*, *333*, 1024–1026.

Diniz-Filho, J. A. F., Rangel, T. F. L. V. B., & Bini, L. M. (2008). Model selection and information theory in geographical ecology. *Global Ecology and Biogeography*, *17*, 479–488.

Dobrowski, S. Z., Thorne, J. H., Greenberg, J. A., Safford, H. D., Mynsberge, A. R., Crimmins, S. M., & Swanson, A. K. (2011). Modeling plant ranges over 75 years of climate change in California, USA: Temporal transferability and species traits. *Ecological Monographs*, *81*, 241–257.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., … Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*, 27–46.

Dormann, C. F., Schymanski, S. J., Cabral, J., Chuine, I., Graham, C., Hartig, F., ... Singer, A. (2012). Correlation and process in species distribution models: Bridging a dichotomy. *Journal of Biogeography*, *39*, 2119–2131.

Edwards, J. L., Lane, M. A., & Nielsen, E. S. (2000). Interoperability of biodiversity databases: Biodiversity information on every desktop. *Science*, *289*, 2312–2314.

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*, 677–697.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17*, 43–47.

Esselman, P. C., & Allan, J. D. (2011). Application of species distribution models and conservation planning software to the design of a reserve network for the riverine fishes of northeastern Mesoamerica. *Freshwater Biology*, *56*, 71–88.

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, *24*, 38–49.

Fourcade, Y., Engler, J. O., Rödder, 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*, e97122.

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I. T., ... Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, *16*, 1424–1435.

Hanley, A., & McNeil, J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36.

Hijmans, R. J. (2012). Cross-validation of species distribution models: Removing spatial sorting bias and calibration with a null model. *Ecology*, *93*, 679–688.

Hijmans, R. J. (2014). *raster: Geographic data analysis and modeling* (R package version 2.2–12). Retrieved from http://CRAN.R-project.org/package=raster

Hijmans, R. J., & Graham, C. H. (2006). The ability of climate envelope models to predict the effect of climate change on species distributions. *Global Change Biology*, *12*, 2272–2281.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*, 1965–1978.

Jiménez-Valverde, A. (2014). Threshold-dependence as a desirable attribute for discrimination assessment: Implications for the evaluation of species distribution models. *Biodiversity and Conservation*, *23*, 369–385.

Jiménez-Valverde, A. (2012). Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography*, *21*, 498–507.

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*, *22*, 508–516.

Jiménez-Valverde, A., Barve, N., Lira-Noriega, A., Maher, S. P., Nakazawa, Y., Papeş, M., ... Peterson, A. T. (2011). Dominant climate influences on North American bird distributions. *Global Ecology and Biogeography*, *20*, 114–118.

Jiménez-Valverde, A., Peterson, A. T., Soberón, J., Overton, J. M., Aragon, P., & Lobo, J. M. (2011). Use of niche models in invasive species risk assessments. *Biological Invasions*, *13*, 2785–2797.

Liu, C., White, M., & Newell, G. (2013). Selecting thresholds for the prediction of species occurrence with presence-only data. *Journal of Biogeography*, *40*, 778–789.

Lobo, J. M., Jiménez-Valverde, A., & Hortal, J. (2010). The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, *33*, 103–114.

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.

Merow, C., Smith, M. J., & Silander, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: What it does, and why inputs and settings matter. *Ecography*, *36*, 1058–1069.

Morales, N. S., Fernandez, I. C., & Baca-Gonzalez, V. (2017). MaxEnt's parameter configuration and small samples: Are we paying attention to recommendations? A systematic review. *PeerJ*, *5*, e3093.

Murray, J. V., Goldizen, A. W., O'Leary, R. A., McAlpine, C. A., Possingham, H. P., & Choy, S. L. (2009). How useful is expert opinion for predicting the distribution of a species within and beyond the region of expertise? A case study using brush-tailed rock-wallabies Petrogale penicillata. *Journal of Applied Ecology*, *46*, 842–851.

Muscarella, R., Galante, P. J., Soley-Guardia, M., Boria, R. A., Kass, J. M., Uriarte, M., & Anderson, R. P. (2014). ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. *Methods in Ecology and Evolution*, *5*, 1198–1205.

Pearson, R. G., & Dawson, T. P. (2003). Predicting the impacts of climate change on the distribution of species: Are bioclimate envelope models useful? *Global Ecology and Biogeography*, *12*, 361–371.

Peterson, A. T., Papeş, M., & Soberón, J. (2008). Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling*, *213*, 63–72.

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., & Araújo, M. B. (2011). *Evaluating model performance and significance* (pp. 150–181). Princeton, NJ: Princeton University Press.

Petitpierre, B., Broennimann, O., Kueffer, C., Daehler, C., & Guisan, A. (2017). Selecting predictors to maximize the transferability of species distribution models: Lessons from cross-continental plant invasions. *Global Ecology and Biogeography*, *26*, 275–287.

Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation. *Ecography*, *31*, 161–175.

Phillips, S. J., & Elith, J. (2010). POC plots: Calibrating species distribution models with presence-only data. *Ecology*, *91*, 2476–2484.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*, 231–259.

Porfirio, L. L., Harris, R. M. B., Lefroy, E. C., Hugh, S., Gould, S. F., Lee, G., ... Mackey, B. (2014). Improving the use of species distribution models in conservation planning and management under climate change. *PLoS ONE*, *9*, e113749-e113749.

R Development Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Radosavljevic, A., & Anderson, R. P. (2014). Making better Maxent models of species distributions: Complexity, overfitting and evaluation. *Journal of Biogeography*, *41*, 629–643.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., ... Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*, 913–929.

Rödder, D., & Lötters, S. (2010). Explanative power of variables used in species distribution modelling: An issue of general model transferability or niche shift in the invasive Greenhouse frog (*Eleutherodactylus planirostris*). *Naturwissenschaften*, *97*, 781–796.

Rödder, D., Schmidtlein, S., Veith, M., & Lötters, S. (2009). Alien invasive slider turtle in unpredicted habitat: A matter of niche shift or of predictors studied? *PLoS ONE*, *4*, e7843.

Smith, A. B. (2013). On evaluating species distribution models with random background sites in place of absences when test presences disproportionately sample suitable habitat. *Diversity and Distributions*, *19*, 867–872.

Soberón, J. (2007). Grinnellian and Eltonian niches and geographic distributions of species. *Ecology Letters*, *10*, 1115–1123.

Stolar, J., & Nielsen, S. E. (2015). Accounting for spatially biased sampling effort in presence-only species distribution modelling. *Diversity and Distributions*, *21*, 595–608.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*, 1285–1293.

Syfert, M. M., Smith, M. J., & Coomes, D. A. (2013). The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. *PLoS ONE*, *8*, e55158.

Synes, N. W., & Osborne, P. E. (2011). Choice of predictor variables as a source of uncertainty in continental-scale species distribution modelling under climate change. *Global Ecology and Biogeography*, *20*, 904–914.

Thomas, C. D. (2010). Climate, climate change and range boundaries. *Diversity and Distributions*, *16*, 488–495.

Varela, S., Lobo, J. M., & Hortal, J. (2011). Using species distribution models in paleobiogeography: A matter of data, predictors and concepts. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *310*, 451–463.

Veloz, S. D. (2009). Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography*, *36*, 2290–2299.

Verbruggen, H., Tyberghein, L., Belton, G. S., Mineur, F., Jueterbock, A., Hoarau, G., … De Clerck, O. (2013). Improving transferability of introduced species' distribution models: New tools to forecast the spread of a highly invasive seaweed. *PLoS ONE*, *8*, e68337.

von Humboldt, A., & Bonpland, A. (1805). *Essai sur la géographie des plantes; accompagné d'un tableau physique des régions équinoxiales*. Paris, France: Levrault, Schoell & Cie.

Wallace, A. R. (1876). *The geographical distribution of animals; with a study of the relations of living and extinct faunas as elucidating the past changes of the Earth's surface*. New York: Harper & Brothers.

Wang, L., & Jackson, D. A. (2014). Shaping up model transferability and generality of species distribution modeling for predicting invasions: Implications from a study on *Bythotrephes longimanus*. *Biological Invasions*, *16*, 2079–2103.

Warren, D. L., & Seifert, S. N. (2011). Ecological niche modeling in Maxent: The importance of model complexity and the performance of model selection criteria. *Ecological Applications*, *21*, 335–342.

Warren, D. L., Cardillo, M., Rosauer, D. F., & Bolnick, D. I. (2014). Mistaking geography for biology: Inferring processes from species distributions. *Trends in Ecology and Evolution*, *29*, 572–580.

Warren, D. L., Glor, R. E., & Turelli, M. (2008). Environmental niche equivalency versus conservatism: Quantitative approaches to niche evolution. *Evolution*, *62*, 2868–2883.

Warren, D. L., Glor, R. E., & Turelli, M. (2010). ENMTools: A toolbox for comparative studies of environmental niche models. *Ecography*, *33*, 607–611.

Wenger, S. J., & Olden, J. D. (2012). Assessing transferability of ecological models: An underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, *3*, 260–267.

Yackulic, C. B., Chandler, R., Zipkin, E. F., Royle, J. A., Nichols, J. D., Campbell Grant, E. H., & Veran, S. (2013). Presence-only modelling using MAXENT: When can we trust the inferences? *Methods in Ecology and Evolution*, *4*, 236–243.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*, 32–35.

Zeng, Y., Low, B. W., & Yeo, D. C. J. (2016). Novel methods to select environmental variables in MaxEnt: A case study using invasive crayfish. *Ecological Modelling*, *341*, 5–13.

BIOSKETCHES

Yoan Fourcade is a postdoctoral researcher at the Swedish University of Agriculture Sciences, mainly interested in conservation biology and biogeography. He aims at understanding how human activity impacts large-scale and long-term processes such as distribution, adaptation and community composition.

Aurélien Besnard is interested in the applied conservation of organisms, especially in agricultural landscapes.

Jean Secondi's research interests encompass conservation biology and sensory ecology, in the perspective of studying how habitat and interspecific interactions shape ecological and evolutionary responses in vertebrates.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Fourcade Y, Besnard AG, Secondi J. Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecol Biogeogr*. 2018;27:245–256. https://doi.org/10.1111/geb.12684