Ecological Modelling xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

The effects of model and data complexity on predictions from species distributions models

David García-Callejas[a,b,*], Miguel B. Araújo[a,b,c,d]

[a] InBIO/CIBIO, University of Évora, Largo dos Colegiais, 7000 Évora, Portugal
[b] Imperial College London, Silwood Park Campus, Buckhurst Road, Ascot, United Kingdom
[c] National Museum of Natural Sciences, Calle Jose Gutierrez Abascal, 2, 28006 Madrid, Spain
[d] Center for Macroecology, Evolution and Climate, Natural History Museum of Denmark, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen, Denmark

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Climate change
Data complexity
Model complexity
Species distributions models
Transferability
Virtual species

ABSTRACT

How complex does a model need to be to provide useful predictions is a matter of continuous debate across environmental sciences. In the species distributions modelling literature, studies have demonstrated that more complex models tend to provide better fits. However, studies have also shown that predictive performance does not always increase with complexity. Testing of species distributions models is challenging because independent data for testing are often lacking, but a more general problem is that model complexity has never been formally described in such studies. Here, we systematically examine predictive performance of models against data and models of varying complexity. We introduce the concept of computational complexity, widely used in theoretical computer sciences, to quantify model complexity. In addition, complexity of species distributional data is characterized by their geometrical properties. Tests involved analysis of models' ability to predict virtual species distributions in the same region and the same time as used for training the models, and to project distributions in different times under climate change. Of the eight species distribution models analyzed five (Random Forest, boosted regression trees, generalized additive models, multivariate adaptive regression splines, MaxEnt) showed similar performance despite differences in computational complexity. The ability of models to forecast distributions under climate change was also not affected by model complexity. In contrast, geometrical characteristics of the data were related to model performance in several ways: complex datasets were consistently more difficult to model, and the complexity of the data was affected by the choice of predictors and the type of data analyzed. Given our definition of complexity, our study contradicts the widely held view that the complexity of species distributions models has significant effects in their predictive ability while findings support for previous observations that the properties of species distributions data and their relationship with the environment are strong predictors of model success.

1. Introduction

Understanding why species distribute as they do is a central problem in ecology. Current methods for studying species distributions often involve statistical or numerical models that relate the distributions of species with layers of environmental information. The use of correlative species distributions models (also known as bioclimatic envelope models, habitat suitability models, and ecological niche models; for definitions of these seemingly related terms see Araújo and Peterson, 2012) is currently the most widespread approach due to their versatility, ease of use, and modest data requirements (e.g., Guisan and Zimmermann, 2000; Elith and Leathwick, 2009). Yet, despite widespread use of these models, the debate as to what is the best modelling approach is far from settled (Araújo and Rahbek, 2006), and predictions from alternative models can be markedly different in the context of spatial (e.g., Randin et al., 2006; Duncan et al., 2009; Heikkinen et al., 2012) and temporal transferability (e.g., Thuiller, 2004; Araújo et al., 2005a; Pearson et al., 2006; Zanini et al., 2009). Previous tests of performance of species distributions models have led to the conclusion that more complex models were generally better than simpler ones (e.g., Segurado and Araújo, 2004; Elith et al., 2006). However, model performance is typically inflated when test data that are not independent from data is used to train the models, such as when

* Corresponding author at: Centre for Ecological Research and Forestry Applications (CREAF), Universitat Autónoma de Barcelona, Bellaterra, Spain.
Tel.: +34 935868388.
E-mail address: david.garcia.callejas@gmail.com (D. García-Callejas).

http://dx.doi.org/10.1016/j.ecolmodel.2015.06.002
0304-3800/

data are randomly split between training and test sets (Araújo et al., 2005b; but see Madon et al., 2013). In the few cases in which models have been tested for transferability using independent data (from another region or another time), no clear relationship between the perceived complexity of the models and their performance was found (Araújo et al., 2005b; Randin et al., 2006; Dobrowski et al., 2011; Heikkinen et al., 2012; Smith et al., 2013).

Models perceived as 'simple' usually have procedures for fitting the data that are easier to grasp, and/or perform fewer and/or simpler operations with the data. In contrast, models perceived as 'complex' involve procedure for fitting the data that are more difficult to comprehend while usually performing a significant number of operations in order to produce the desired outcome. It is implicitly assumed that this loose definition of complexity is related to the capacity of different models to produce either 'simple' or 'complex' response curves (e.g., Elith et al., 2006; Merow et al., 2014). When selecting the best model for a given problem, it is expected that parsimony should lead to selecting models that minimize overall prediction error by finding an optimal balance between the error in fitting the training data (also referred to as parameter estimation error or bias) and the error in generalizing to new datasets (also referred to as approximation error, or variance). Models that are too simple would fit training data poorly (high bias), while overly complex models would generate low bias and high variance as they would capture random error or biases in the data.

The concept of model complexity is central to the endeavour of finding optimal models for predictive purposes. Yet, measuring model complexity is not straightforward. Here, we attempt to formalize one of the key aspects of model complexity (algorithmic or computational complexity, see Section 1.1: computational complexity), and test whether the principle of parsimony can guide identification of the optimal model complexity for predicting species distributions in space and time. In addition, we quantify structural complexity in the response data (i.e., presence and absence of species) and examine how data complexity affects the predictive abilities of models. To overcome problems of data availability, we simulated virtual species across a realistic geographical domain with different sets of environmental predictors, and compare some of the results with empirical presence–absence records within the same geographical domain.

1.1. Computational complexity

The computational complexity of an algorithm is defined by the amount of computational resources it requires to produce an output (Arora and Barak, 2009). This definition stems from the idea that an algorithm processes an input via a certain number of elementary operations, and these operations consume varying amounts of computing time. The computing time spent by the algorithm is, thus, an approximation to the complexity of the operations performed on the input. Complex algorithms inevitably perform more complex operations on their input than simple ones, thereby requiring more computation time to solve a particular task. Numerical analyses of computational complexity treat algorithms as black boxes, disregarding their internal structure, functional form or any other specificity. Such analyses are, therefore, suitable when the goal is to compare different methodologies on equal footing.

Computational complexity is also referred to as time complexity or algorithmic complexity, and it is commonly expressed by the O notation (read 'big o'). This notation identifies the time complexity of an algorithm by the highest-order term of its growth rate as a function of input size, suppressing lower-order terms and constants. It is an asymptotic measure of complexity; as input size increases, so does the importance of the dominant term in characterizing computation time. For example, an algorithm may take $3x^2 + 2x$ time units in solving a problem of size x. As x approaches infinity, the higher-order term (x^2) will tend to take over computation time, and the lower-order terms, as well as the multiplicative coefficients, will become irrelevant. This particular algorithm is thus said to have a time complexity of $O(n^2)$. If the computation time of an algorithm is independent of the dataset size, it is said to be a constant time algorithm, expressed as a time complexity of $O(1)$. As this methodology aims to estimate the asymptotical behaviour of a given algorithm, it also bypasses the issue of comparing algorithms written in different programming languages: the computational cost of a given algorithm implemented in two different languages is assumed to be proportional, up to a multiplicative constant that will become irrelevant asymptotically. The chief assumption of the method is that the algorithms being compared are efficiently programmed, i.e., there are no spurious tasks within the algorithms consuming computation time. A full treatment of computational complexity is out of the scope of this study (but see Arora and Barak, 2009; Papadimitriou, 1994).

Consistent with the principle of parsimony and assuming that algorithmic complexity is a good proxy for overall model complexity, the highest predictive capacity should be expected in models of intermediate algorithmic complexity. It is worth noting that the quantification of algorithmic complexity is independent of the modelling methodology. That is, the framework implemented herein with correlative species distributions models, could also be easily implemented with alternative mechanistic approaches for modelling species distributions (e.g., Fordham et al., 2013; García-Valdés et al., 2015).

1.2. Geometrical complexity of the data

Estimating the ecological niche of a species is an instance of the broad class of problems in which a set of points (in environmental space) must be classified into one of two opposing classes (presence/absence) according to some relationship between the dimensions of the space and the class to which each point belongs. The difficulty in estimating the ecological niche of a species can be assessed by evaluating the geometrical structure of the boundary between classes in the training data. Aside from deficiencies and biases in the data collection (Barry and Elith, 2006; Araújo et al., 2009), the internal structure of the data and its relationship with models predictive capacity has never been formally explored (but see Blonder et al., 2014). It has been, though, extensively addressed in other scientific fields; particularly within the machine learning community where the concept of *geometrical complexity* has been developed. Given a dataset with a two-class categorical response and N predictors, the geometrical complexity is defined as an approximation of the structural characteristics of the N-dimensional boundary separating the response classes (Basu and Ho, 2006). It is a general measure defined by a set of complementary metrics (see Section 2). When analyzed together, these metrics help differentiate datasets with geometrically simple class boundaries from those with complex and/or random class boundaries.

We predict that data complexity is related to the predictive capacity of the models. Specifically, simpler datasets will tend to reflect simpler occurrence–environment relationships thus being easier to model and yielding comparatively better performance than models trained with more complex datasets.

2. Materials and methods

2.1. Virtual species generation

We created three different types of virtual species. Their distributions were projected across mainland Spain by defining environmental suitability landscapes based on different sets of

Please cite this article in press as: García-Callejas, D., Araújo, M.B., The effects of model and data complexity on predictions from species distributions models. Ecol. Model. (2015), http://dx.doi.org/10.1016/j.ecolmodel.2015.06.002

Table 1
Variables used in the creation of the three types of virtual species.

Species type	Variables used
Type I (annual response)	Mean annual temperature, mean annual precipitation
Type II (seasonal response)	Mean annual temperature, mean annual precipitation, precipitation seasonality
Type III (spatial response)	Mean annual temperature, mean annual precipitation, vegetation basal area of each cell

environmental covariates resampled within a 1 km × 1 km grid (Table 1 and Appendix A). Following Valladares et al. (2014), we assumed that the suitability of the environment for species followed nonlinear functional forms, and the overall environmental suitability was the product of the partial suitabilities for each environmental correlate. The first type of virtual species, labelled 'annual', comprised nine species with different combinations of Gaussian and Beta response curves for mean annual temperature and mean annual precipitation (Appendix A). For defining the second type of virtual species ('spatial'), we added, in addition to mean annual temperature and precipitation, a spatially explicit covariate (Table 1). The spatial variable was basal area of forest species for each cell, as measured by the Third National Forest Inventory of Spain (IFN, Ministerio de Medio Ambiente, 2006). For the third type of virtual species, labelled 'seasonal', in addition to mean annual temperature and precipitation, we constrained the potential suitability with an index of precipitation seasonality obtained from the WorldClim database (Hijmans et al., 2005). Two 'spatial' species and two 'seasonal' were created, with varying optima for each covariate. From the set of 9 annual, 2 spatial, and 2 seasonal continuous response curves, we derived 26 more suitability distributions by adding two white noise filters to each surface, varying the overall suitability in 0.25*e and 0.5*e, where $e \sim N(0,sd(suitability))$. A fixed threshold of 0.5 was further applied to define presences and absences. The combined use of the selected stochastic filters and the fixed threshold effectively blurs the separation between presences and absences at geographic range limits, without altering significantly the outcome for areas with very high or low suitability (Appendix A). Furthermore, the use of a variable presence threshold for generating virtual distributions is analyzed in Appendix C.

Using the response curves of each species, we then projected species presence–absence maps in 2090, using the CGCM2 Global Climate Model under the A2 climate scenario (IPCC, 2007). While mean annual temperature, mean annual precipitation, and the seasonality index in precipitation were taken from the climate model, the basal area of vegetation was assumed constant to the levels of 2010. We also assumed no limitations to dispersal for the projected virtual species.

For comparison with our virtual species distributions, we also modelled the distributions of 16 empirical tree species across mainland Spain. These data were again obtained from the third national forest inventory of Spain (Ministerio de Medio Ambiente, 2006). All Spanish territory was consistently sampled in a grid of 1 km × 1 km for the inventory. Consequently, it is reasonable to assume true absences in this dataset.

2.2. Modelling methods and projections

Eight methods were used to model species distributions, considering three broad modelling approaches. Specifically, a surface-range envelope method: BIOCLIM (Booth et al., 2014); three regression-based methods: GLM with first order terms (Guisan and Zimmermann, 2000), GAM (Hastie and Tibshirani, 1990) and MARS (Friedman, 1991); and four machine learning algorithms: MaxEnt (Elith et al., 2011), boosted regression trees (BRT, Elith et al., 2008), Random Forest (Breiman, 2001) and support vector machines (SVM, Vapnik, 1998) with a linear kernel. Algorithms were applied using the base R implementation for GLM, and the packages *dismo* (for BIOCLIM and the interface to MaxEnt), *gbm* (for BRT), *earth* (for MARS), *mgcv* (for GAM), *randomForest*, and *e1071* (for SVM) in R 3.0.3 (R Core Team, 2014). Throughout the simulations, model-specific parameters were kept constant at recommended values found in the literature (Appendix B).

We undertook two types of simulations: the first for testing the ability of models to characterize the baseline distributions of the virtual and empirical species in 2010; and the second for testing the ability of the models calibrated in 2010 to predict distributions of the virtual species in 2090.

In the first set of simulations, for each species, 10 training sets were sampled each consisting of 200 presences and 200 true absences. We selected training sets without sampling error or bias, so as not to include more confounding factors in our analyses. Metrics of geometrical complexity of data for training sets were obtained and averaged, and model fits were averaged and evaluated against the known distributions. In the second set of simulations, models were trained and results averaged again over 10 training sets of 200 presences and 200 true absences. Data complexity metrics were obtained, and models were evaluated against the 2090 known (virtual) projections. In model projections, we used as predictors the two climate variables common to all virtual species (Table 1), deliberately omitting the other variables used to simulate presence and absence of the spatial and seasonal species (we also conducted simulations with the full set of environmental variables as predictors, see Appendix C). This allowed us to weigh the influence of omitting different types of covariates in models predictive capacity. Model performance was evaluated using the area under the receiving operator curve (AUC). This performance metric is particularly useful when comparing models including true absences (e.g., Jiménez-Valverde, 2012). Trends in AUC scores were also compared with those of another accuracy metric recently proposed, Se^* (Jiménez-Valverde, 2014; see Appendix C).

Variation in AUC scores was analyzed with linear mixed-effects models, taking AUC as response and the following fixed effects: model, type of simulation (no transferability or temporal transferability) and type of species. Species were included as a random effect nested within species type. We used the Kenward-Roger approximation for obtaining p-values associated to the significance of the fixed effects (Halekoh and Højsgaard, 2014), and evaluated the inclusion of the random effect by comparing the AIC value of models with and without the species random effect. Analyses were performed using the *lme4* and *pbkrtest* packages in R (R Core Team, 2014).

2.3. Computational complexity

The time complexity of each technique was numerically approximated by measuring the computation time for training models with datasets of increasing number of records. The exact expression of the time complexity of a given algorithm is virtually impossible to obtain with numerical approximations, but the broad asymptotical relationship between computation time and dataset size can be inferred. An analytical expression of the time complexity, on the other hand, can be obtained only after a detailed evaluation of the algorithm's source code. Such evaluation is impractical in most cases, especially when comparing a significant number of algorithms comprising several hundreds or thousands of lines of code. Computation time was measured with the R function *proc.time*. Models were fitted for datasets with two predictor variables and a categorical response (i.e., species distributions) ranging from 100 000 to 750 000 observations, with every other parameter in each model kept constant. As we were not interested in the results of the model fitting, but only on the computation

Please cite this article in press as: García-Callejas, D., Araújo, M.B., The effects of model and data complexity on predictions from species distributions models. Ecol. Model. (2015), http://dx.doi.org/10.1016/j.ecolmodel.2015.06.002

time, we used randomly generated data of the desired length. Datasets of less than 100 000 observations showed no clear trend for most models and a high number of nil values for computation time. Each measurement reported is the average of 50 bootstrap repetitions in the same computer (intel i7 processor with 16 GB RAM) under the same conditions. For analyzing the asymptotical behaviour of each model, we estimated the type of curve that best described the relationship between dataset size and computation time, considering the broad categories of linear and polynomial relationships (no model appeared to exhibit an exponential behaviour). A linear relationship is recovered by a linear model on the untransformed data, and a polynomial relationship $y = \alpha \times x^{\beta}$ is recovered by a linear model on the log–log transformed data: $\log(y) \sim \log(\alpha \times x^{\beta}) = \log(\alpha) + \beta \times \log(x)$. We fitted these alternatives to the computation times obtained for each species distribution modelling technique, and assumed that the correct relationship was given by the data that best approximated a linear correlation, measured by the Pearson product-moment correlation coefficient (PCC). Models falling within the same complexity category were sorted according to the slope of their regression curves.

2.4. Data complexity

The overarching concept of geometrical complexity has proved to be virtually impossible to capture with a single metric, because there are several aspects of the structure of the boundary between categories that need to be accounted for (Basu and Ho, 2006). From the range of metrics developed in the studies that first proposed the geometrical complexity concept (Ho and Basu, 2002; Basu and Ho, 2006; Ho, 2008), our own preliminary analyses showed that two of them were especially effective in discriminating our pool of virtual and empirical datasets. These are the *ratio of intra/interclass nearest neighbour distance* and the *nonlinearity of a classifier*.

2.4.1. Ratio of intra/interclass nearest neighbour distance

Ratio of intra/interclass nearest neighbour distance is a measure of the spread within classes relative to the spread among classes. For each record, the distance to the nearest neighbour of its class and the distance to the nearest neighbour of its opposite class are recorded. These quantities are averaged over the entire dataset, and their ratio is the metric. The metric is sensitive to the magnitude of the dispersion in environmental space within classes relative to the environmental gap between classes: high values of the metric are less exact than low values.

2.4.2. Nonlinearity of a classifier

The nonlinearity rate for a given dataset is defined as the probability that an arbitrary point, uniformly and linearly interpolated between two arbitrary points in the dataset with the same classification, shares this classification (Hoekstra and Duin, 1996). It is intended to give a measure of the nonlinearity of the N-dimensional boundary between classes. First, a test set is created by interpolating randomly sampled points from the same class. Then, the classifier is applied, and the error rate of the classifier over the interpolated set is the metric. With the implementation proposed here, using a nearest-neighbour classifier, a high error rate implies that many interpolated points do not share the classification of their nearest neighbours, implying a high nonlinearity of the classification boundary. The metric is, therefore, sensitive to the geometry of the class boundary.

All in all, both metrics measure complementary aspects of the geometrical complexity of a classification problem. Other metrics implemented are reviewed in Appendix D. In Fig. 1, the most complex situation is that of panel (d), but different dataset structures give rise to different combinations of the metrics. Geometrical complexity metrics are built upon the N-dimensional distances between points of the environmental space. In our analyses, these distances were computed using a variation of the N-dimensional Euclidean distance, called the Heterogeneous Value Distance Metric that is more accurate than the former and able to generalize over categorical and numeric covariates seamlessly (Wilson and Martinez, 1997).

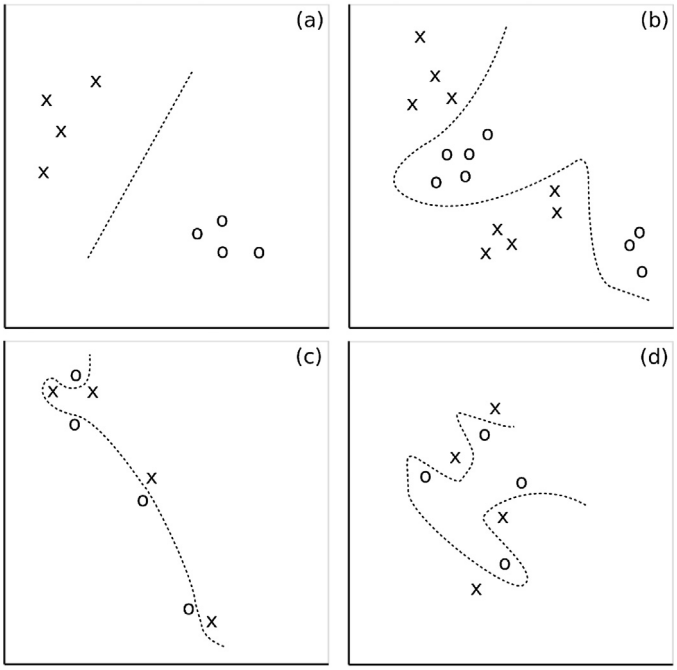

Fig. 1. Data complexity as measured with different geometrical complexity metrics. A two-dimensional environmental space with two response classes is defined where a set of presence and absence samples are plotted. Shown, data with (a) low ratio intra/interclass distance and low nonlinearity; (b) low ratio intra/interclass distance and high nonlinearity; (c) high ratio intra/interclass distance and low nonlinearity; (d) high ratio intra/interclass distance and high nonlinearity. Dashed lines represent potential class boundaries.

3. Results

3.1. Computational complexity

The asymptotical behaviour of all techniques was best approximated by a linear relationship between the dataset size and computation time, but the slope of the linear regression differed several orders of magnitude among the algorithms used to model species distributions (Table 2). The only exception was BIOCLIM that proved insensitive to the size of the dataset and performed equally fast, on average, for any potential number of data points.

Table 2
Pearson product-moment correlation coefficient (PCC) between computation time and dataset size for the original data (indicating a linear relationship) and the log–log transformed data (indicating a polynomial relationship). Bold results indicate the highest PCC value. Also shown the slope of the linear regression that generates the best fit for each model.

Model	Linear PCC	Polynomial PCC	Slope of best fit
BRT	**0.921**	0.817	2.229e−7
GAM	**0.966**	0.833	1.669e−6
GLM	**0.838**	0.758	5.951e−9
MARS	**0.912**	0.826	2.695e−8
MaxEnt	**0.941**	0.938	2.176e−7
Random Forest	**0.996**	0.996	2.399e−5
SVM	**0.97**	0.851	1.191e−6

Please cite this article in press as: García-Callejas, D., Araújo, M.B., The effects of model and data complexity on predictions from species distributions models. Ecol. Model. (2015), http://dx.doi.org/10.1016/j.ecolmodel.2015.06.002

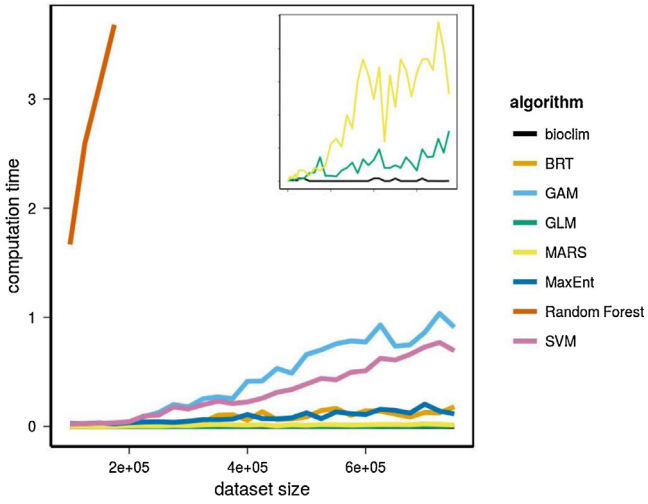

Fig. 2. Computation time for different SDM algorithms and dataset sizes, ranging from 100 000 to 750 000 points (except for Random Forest that became computationally too expensive for sizes of more than 200 000 points). The inset shows the same plot at a smaller scale for computation time, to showcase the different behaviour of MARS, GLM and BIOCLIM.

Among the other methodologies, the Random Forest algorithm displayed a clearly distinct pattern from the remaining techniques, becoming too computationally expensive for datasets of more than 2×10^5 points (Fig. 2). The other methodologies only started to differentiate for datasets bigger than 2×10^5 points. SVM and GAM had similar behaviour, and these in turn were more computationally expensive than BRT and MaxEnt. MARS and GLM were the most efficient algorithms, with the said exception of BIOCLIM.

3.2. Model accuracy and transferability

AUC scores were significantly influenced by model technique (Kenward-Roger corrected test: $F = 61.06$, df = 7, p-value < 0.0001), and type of simulation (Kenward-Roger corrected test: $F = 839.18$, df = 1, p-value < 0.0001). The random effect of species, nested within species type, significantly improved the AIC of the model, and was thus maintained for the analyses. For the simulations that did not involve transferability, the highest mean AUC scores were obtained by MaxEnt (0.97 ± 0.02) and GAM (0.97 ± 0.03), and all methodologies ranged from fair to excellent AUC scores, according to the Swets criteria ($0.5 \leq$ AUC < 0.6 = fail; $0.6 \leq$ AUC < 0.7 = poor; $0.7 \leq$ AUC < 0.8 = fair; $0.8 \leq$ AUC < 0.9 = good; $0.9 \leq$ AUC = excellent; Swets, 1988). In simulations involving temporal transferability, performance was consistently lower and more variable for all type of species and modelling techniques (Figs. 3 and 4). Particularly important was the decrease in predictive capacity for the seasonal species, due to the strong variation of the precipitation seasonality index in the climate scenario used for generating the virtual species (from 48.25 ± 15.26 in 2010 to 70 ± 20.25 in 2090). With temporal transferability, models with the best performance were MaxEnt (0.82 ± 0.18), MARS (0.81 ± 0.19) and Random Forest (0.81 ± 0.20). GLM and SVM consistently performed worse than the other methodologies, regardless of the type of species modelled (Fig. 3) or the type of simulation (Fig. 4). Additional simulations in which we accounted for the environmental predictors deliberately omitted from the model projections improved significantly the performance of all models, particularly for the two seasonal species in the temporal transferability set of simulations (Appendix C: Tables C.1–C.4).

3.3. Data complexity of the training sets

Virtual datasets obtained consistently lower values for the two geometrical complexity metrics of data complexity used herein (see Section 2) than species from the IFN dataset did (Fig. 5). Preliminary analyses showed that the addition of stochasticity in the data significantly increased both the nonlinearity and the ratio of intra/interclass distance metrics for all types of virtual species (Appendix C: Table C.7, Figs. C.4 and C.5). Those datasets with a clear delimitation between presences and absences (low nonlinearity) and with low ratio of intra/interclass distance obtained higher AUC values (Fig. 5). AUC scores were significantly correlated with the two metrics of data complexity in the simulations not involving transferability for all models except SVM and GLM (Appendix C: Table C.5). In the temporal transferability simulations, significant correlations were found between the nonlinearity metric of data complexity and AUC score for MARS, MaxEnt, BRT and Random Forest, but not for other models or the ratio of intra/interclass distance metric of data complexity (Appendix C: Table C.6).

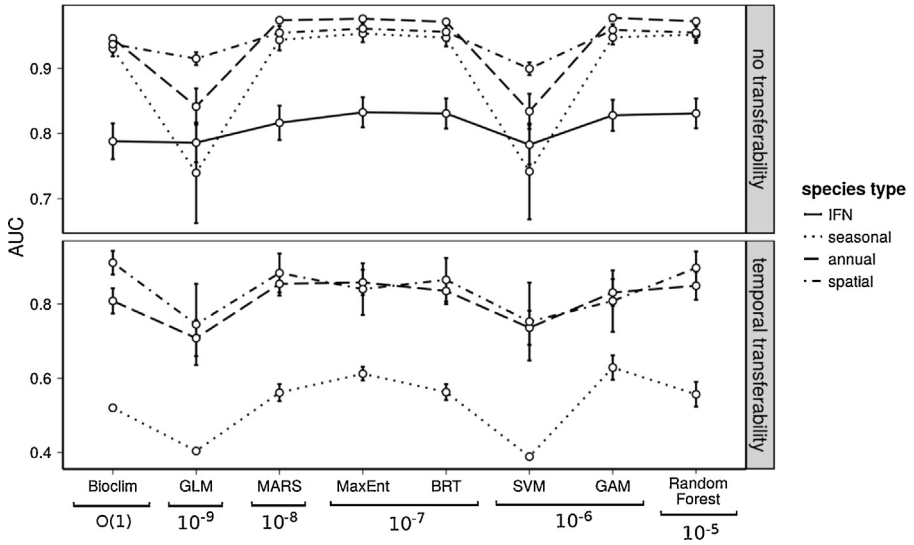

Fig. 3. AUC values of each algorithm for two simulations. (a) Using subsets of the same data for training and testing; and (b) using projections to 2090 to evaluate transferability in time. Error bars represent standard errors. Groupings in the horizontal axis represent complexity categories: BIOCLIM is a constant time model O(1) and the others are O(n) models, sorted by the magnitude of the slope of the linear regression between dataset size and computation time (Table 2).

Please cite this article in press as: García-Callejas, D., Araújo, M.B., The effects of model and data complexity on predictions from species distributions models. Ecol. Model. (2015), http://dx.doi.org/10.1016/j.ecolmodel.2015.06.002

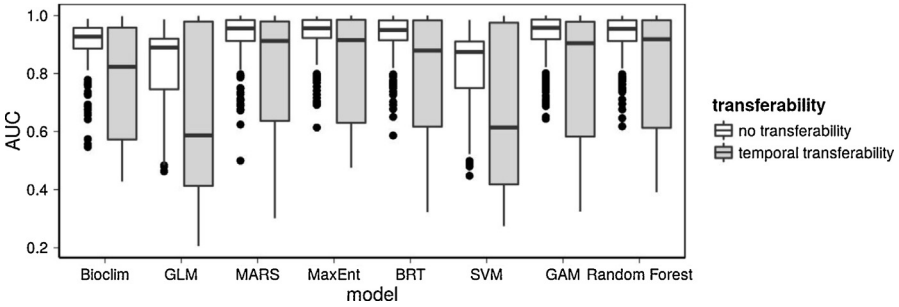

Fig. 4. Distribution of AUC scores for the two types of experiments (no transferability and temporal transferability), evaluated for all models and averaged over all virtual species.

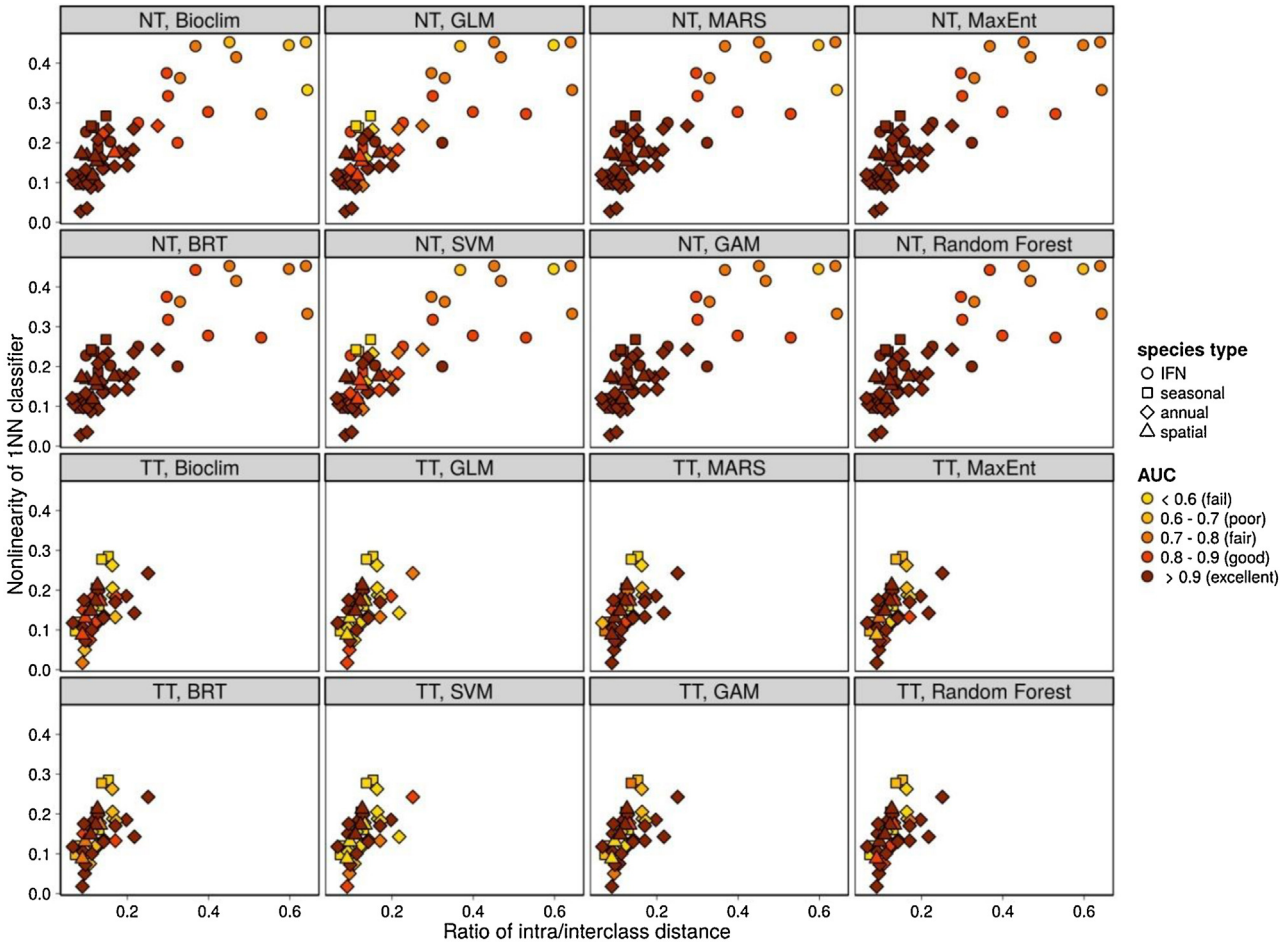

Fig. 5. Two aspects of geometrical complexity (nonlinearity of a 1-nearest neighbour classifier and ratio of intra/interclass distance) and AUC of the models considered. Points represent the mean AUC and geometrical complexity obtained for each species, either virtual or empirical. NT, no transferability simulations; TT, temporal transferability simulations. Note how the IFN datasets are not represented in the temporal transferability simulations.

4. Discussion

We asked how computational complexity of species distributions models and the geometrical complexity of the species distributions data affected the performance of species distributions modelling techniques with and without temporal transferability. The starting assumption was that models of intermediate complexity would tend to show increased performance, while simpler data would be easier to model. Consistent with our hypothesis, we found that data complexity was inversely related with model performance (with and without transferability), whereas the computational complexity of the models was not related to model performance.

4.1. Is computational complexity a reliable estimator of predictive capacity?

With the exception of BIOCLIM, all models fitted fall in the broad category of linear time algorithms, i.e., algorithms that show a linear relationship between computation time and input size. BIOCLIM is a remarkable exception due to its fast computation, independent from dataset size. Although its performance is significantly lower than MARS, MaxEnt, BRT, GAM or Random Forest, it also outperforms GLM and SVM in our model configurations and data sets. Note, however, that for design clarity, GLM were fitted allowing only first-order terms and no interactions, and SVM were fitted with a linear kernel, effectively forcing these two methodologies to

Please cite this article in press as: García-Callejas, D., Araújo, M.B., The effects of model and data complexity on predictions from species distributions models. Ecol. Model. (2015), http://dx.doi.org/10.1016/j.ecolmodel.2015.06.002

model linear responses to the covariates. More generally, we found no relationship between model computational complexity and performance (Fig. 4). Random Forest was 1 order of magnitude slower than the next models, with no significant increase in performance on any simulation. We suggest that when computational cost is an issue, Random Forest might well be avoided. One of the objectives of evaluating models in two sets of simulations was to look for overfitting, specially in the most complex models: if more complex models overfit the training data, we would expect to find decreasing AUC scores the temporal transferability simulations. We found no evidence for this hypothesis, and instead found that all models showed the same pattern of decreasing AUC scores regardless of their computational complexity (Fig. 4). In other words, either the computational complexity of species distributions models is not good a proxy for overall model complexity or the complexity of species distributions models is unrelated to model performance, both in the contexts of no transferability and temporal transferability.

Clearly, other unaccounted factors can affect computational complexity: the number of covariates directly influences the computational cost of certain models (Cutler et al., 2007), as well as the complexity of the functional response obtained; model-specific parameters in machine learning methods can also potentially alter the computational complexity curves obtained (Elith et al., 2008). But it is also possible that models exploring the relationship between species distributions and environmental covariates are powerful at explaining a given pattern but are poor predictors of future patterns given that mechanisms driving distribution are not explicitly accounted for; if the estimated relationships between distributions and environmental covariates are indirect (sensu Austin, 2002), then model complexity is bound to be unrelated to predictive capacity of models for transferability.

4.2. Does geometrical complexity of the data affect model performance?

Previous studies showed that the structural characteristics of data influenced projections of species distribution models (e.g., Segurado and Araújo, 2004; Brotons et al., 2004; Lobo, 2008; Foody, 2011; Moudrý and Šímová, 2012). We have quantified the complexity of distributional data with novel metrics, drawn from the machine learning field, and found that data complexity is significantly related to model performance. Several insights can be outlined from these results. The empirical datasets analyzed here spanned a high range of values for the two complexity metrics used, but most of these datasets showed higher values than the virtual data generated and, consequently, lower AUC scores (Fig. 5). The appearance of high complexity for some datasets associated with complex or quasi-random structural characteristics may indicate (1) the existence of important covariates missing in the experiment, (2) a highly complex occurrence–environment relationship, or (3) the existence of other sources of variability. Note that only the second point relates to the intrinsic complexity of the response of the taxa to the environmental covariates, while the other two have extrinsic causes that relate to deficiencies in sampling or experimental design. An intrinsically complex occurrence–environment relationship can reflect, among other patterns, a high variability in individual responses to environmental factors, or if life stages are aggregated in the presence count, it may arise from differences in the occurrence–environment response among life stages. Concerning extrinsic factors, minimizing these external sources of variability will lead to better performance with any given model (Barry and Elith, 2006) by decreasing the geometrical complexity of the data. For example, an adequate sampling across all dimensions of the environmental space may decrease the ratio of intra/interclass distance by spanning the records of presences and absences over bigger regions of the environmental space. The nonlinearity metric is potentially affected by the accuracy of the presence/absence records, i.e., the ratio of false positives and false negatives in the sample.

The data complexity metrics presented here can help design of virtual datasets better approaching the structural characteristics of empirical data. The key factors influencing the complexity of empirical data (i.e., omission of important covariates, complex occurrence–environment relationships, and existence of other sources of variability) can be manipulated for generating and analyzing virtual datasets. In our study, the omission of covariates when projecting virtual species significantly altered the geometrical complexity of the datasets and, accordingly, decreased model performance compared to data with the full set of covariates (Appendix C: Tables C.1–C.16 and Fig. C.5). The addition of white noise to the virtual datasets also increased significantly both complexity metrics (Appendix C: Table C.19, Figs. C.8 and C.9). Either option made our virtual data closer to the complexity levels of the empirical datasets. However, randomization of the data has the drawback that the causal relationship between covariates and distribution may become too blurred, thus preventing any useful analyses on the virtual data (e.g., see Appendix A). In order to keep as much control as possible over the sources of variability, the use of complex response functions is often preferable, including mechanistic models based in population (Brook et al., 2009; Pagel and Schurr, 2012) or individual-level processes (Matias et al., 2014) and, if necessary, the deliberate omission of covariates for model evaluation.

We have not conducted analyses about the effects of other features sources of data complexity (e.g., grain size, systematic biases). To the extent that different ecological processes influence distribution and abundance of taxa at different scales, datasets at different scales will showcase different degrees of geometrical complexity, representative of the differential role of each ecological process. Datasets at biogeographical scales may, for instance, yield comparatively simple geometrical complexity due to the averaging out of local ecological processes and the strong large-scale signal of environmental forcing (Pearson and Dawson, 2003). Therefore, while the comparison of geometrical complexity of categorical datasets may potentially be useful for different types of analyses, caution should be taken to verify potential inconsistencies between the different datasets.

5. Conclusions

We evaluated different aspects of complexity related to (1) the computational cost of eight SDM algorithms, and (2) the geometric characteristics of species distributions data. MARS, MaxEnt, BRT and GAM fared equally well as Random Forest with much less computational costs while BIOCLIM performed worse than these five methods but better than GLM and SVM, which were the worst-performing methods with and without temporal transferability. Consistent with previous studies, the capacity of models to predict events in the future under climate change, i.e., to transfer in time, were significantly reduced when compared with their ability to characterize the training data in the baseline period. However, loss of predictive ability of the models when used to transfer species distributions in time was independent of their computational complexity, thus failing to support the original hypothesis that models of intermediate complexity would have greater performance.

In contrast to model complexity, the geometrical complexity of the data was shown to be a strong predictor of model performance. For most modelling methodologies, both with and without temporal transferability, there was a statistically significant negative correlation between predictive performance and geometrical complexity metrics.

Please cite this article in press as: García-Callejas, D., Araújo, M.B., The effects of model and data complexity on predictions from species distributions models. Ecol. Model. (2015), http://dx.doi.org/10.1016/j.ecolmodel.2015.06.002

Acknowledgements

We thank Babak Naimi, Francisco Ferri-Yáñez, Manuel Mendoza, Alejandro Rozenfeld and an anonymous reviewer for discussion. This study was funded through the Integrated Program of IC&DT Call No 1/SAESCTN/ALENT-07-0224-FEDER-001755. D.G.-C. acknowledges additional support from the Spanish Ministry of Education (FPU fellowship).

Appendix A, B, C, D. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ecolmodel.2015.06.002

References

Araújo, M.B., Peterson, A.T., 2012. Uses and misuses of bioclimatic envelope modeling. Ecology 93 (7), 1527–1539, http://dx.doi.org/10.1890/11-1930.1
Araújo, M.B., Rahbek, C., 2006. How does climate change affect biodiversity? Science 313, 1396–1397, http://dx.doi.org/10.1126/science.1131758
Araújo, M.B., Whittaker, R.J., Ladle, R.J., Erhard, M., 2005a. Reducing uncertainty in projections of extinction risk from climate change. Glob. Ecol. Biogeogr. 14, 529–538, http://dx.doi.org/10.1111/j.1466-822x.2005.00182.x
Araújo, M.B., Pearson, R.G., Thuiller, W., Erhard, M., 2005b. Validation of species–climate impact models under climate change. Glob. Change Biol. 11, 1504–1513, http://dx.doi.org/10.1111/j.1365-2486.2005.01000.x
Araújo, M.B., Thuiller, W., Yoccoz, N.G., 2009. Reopening the climate envelope reveals macroscale associations with climate in European birds. Proc. Natl. Acad. Sci. 106, 45–46.
Arora, S., Barak, B., 2009. Computational Complexity: A Modern Approach. Cambridge University Press, http://dx.doi.org/10.1017/CBO9780511804090
Austin, M., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecol. Model. 157, 101–118, http://dx.doi.org/10.1016/S0304-3800(02)00205-3
Barry, S., Elith, J., 2006. Error and uncertainty in habitat models. J. Appl. Ecol. 43 (3), 413–423, http://dx.doi.org/10.1111/j.1365-2664.2006.01136.x
Basu, M., Ho, T.K. (Eds.), 2006. Data Complexity in Pattern Recognition. Springer-Verlag, London, http://dx.doi.org/10.1007/978-1-84628-172-3
Blonder, B., Lamanna, C., Violle, C., Enquist, B.J., 2014. The n-dimensional hypervolume. Glob. Ecol. Biogeogr. 23, 595–609, http://dx.doi.org/10.1111/geb.12146
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. Divers. Distrib. 20, 1–9, http://dx.doi.org/10.1111/ddi.12144
Breiman, L., 2001. Random forests. Machine Learning, 45., pp. 5–32, http://dx.doi.org/10.1023/A:1010933404324
Brook, B.W., Akçakaya, H.R., Keith, D.A., Mace, G.M., Pearson, R.G., Araújo, M.B., 2009. Integrating bioclimate with population models to improve forecasts of species extinctions under climate change. Biol. Lett. 5 (6), 723–725, http://dx.doi.org/10.1098/rsbl.2009.0480
Brotons, L., Thuiller, W., Araújo, M.B., Hirzel, A.H., 2004. Presence–absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography 27, 437–448, http://dx.doi.org/10.1111/j.0906-7590.2004.03764.x
Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88 (11), 2783–2792, http://dx.doi.org/10.1890/07-0539.1
Dobrowski, S.Z., Thorne, J.H., Greenberg, J.A., Safford, H.D., Mynsberge, A.R., Crimmins, S.M., et al., 2011. Modeling plant ranges over 75 years of climate change in California USA: temporal transferability and species traits. Ecol. Monogr. 81, 241–257, http://dx.doi.org/10.1890/10-1325.1
Duncan, R.P., Cassey, P., Blackburn, T.M., 2009. Do climate envelope models transfer? A manipulative test using dung beetle introductions. Proc. R. Soc. B: Biol. Sci. 276, 1449–1457, http://dx.doi.org/10.1098/rspb.2008.1801
Elith, J., Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40 (1), 677–697, http://dx.doi.org/10.1146/annurev.ecolsys.110308.120159
Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Mcc Overton, J., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S., Zimmermann, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151, http://dx.doi.org/10.1111/j.2006.0906-7590.04596.x
Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77 (4), 802–813, http://dx.doi.org/10.1111/j.1365-2656.2008.01390.x
Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J., 2011. A statistical explanation of MaxEnt for ecologists. Divers. Distrib. 17, 43–57, http://dx.doi.org/10.1111/j.1472-4642.2010.00725.x
Foody, G.M., 2011. Impacts of imperfect reference data on the apparent accuracy of species presence–absence models and their predictions. Glob. Ecol. Biogeogr. 20, 498–508, http://dx.doi.org/10.1111/j.1466-8238.2010.00605.x
Fordham, D.A., Akçakaya, H.R., Araújo, M.B., Keith, D.A., Brook, B.W., 2013. Tools for integrating range change, extinction risk and climate change information into conservation management. Ecography 36, 956–964, http://dx.doi.org/10.1111/j.1600-0587.2013.00147.x
Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Stat. 19, 1–67.
García-Valdés, R., Gotelli, N.J., Zavala, M.A., Purves, D., Araújo, M.B., 2015. Effects of climate, species interactions, and dispersal on decadal colonization and extinction rates of Iberian tree species. Ecol. Model. 309, 118–127.
Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135 (2–3), 147–186, http://dx.doi.org/10.1016/S0304-3800(00)00354-9
Halekoh, U., Højsgaard, S., 2014. A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models – the R package pbkrtest. J. Stat. Softw. 59 (9), 1–30.
Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman and Hall.
Heikkinen, R.K., Marmion, M., Luoto, M., 2012. Does the interpolation accuracy of species distribution models come at the expense of transferability? Ecography 35 (3), 276–288, http://dx.doi.org/10.1111/j.1600-0587.2011.06999.x
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. Int. J. Climatol. 25, 1965–1978, http://dx.doi.org/10.1002/joc.1276
Ho, T.K., 2008. Data complexity analysis: linkage between context and solution in classification. In: Structural Syntactic, and Statistical Pattern Recognition. Springer, Berlin, pp. 986–995.
Ho, T.K., Basu, M., 2002. Complexity measures of supervised classification problems. IEEE Trans. Pattern Anal. Mach. Intel. 24, 289–300.
Hoekstra, A., Duin, R.P.W., 1996. On the nonlinearity of pattern classifiers. In: Proc. of the 13th ICPR, pp. 271–275.
IPCC., 2007. Climate Change Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC, Geneva, Switzerland.
Jiménez-Valverde, A., 2012. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Glob. Ecol. Biogeogr. 21 (4), 498–507, http://dx.doi.org/10.1111/j.1466-8238.2011.00683.x
Jiménez-Valverde, A., 2014. Threshold-dependence as a desirable attribute for discrimination assessment: implications for the evaluation of species distribution models. Biodivers. Conserv. 23, 369–385, http://dx.doi.org/10.1007/s10531-013-0606-1
Lobo, J.M., 2008. More complex distribution models or more representative data? Biodivers. Inform. 82, 14–19.
Madon, B., Warton, D.I., Araújo, M.B., 2013. Community-level vs species-specific approaches to model selection. Ecography 36, 1291–1298, http://dx.doi.org/10.1111/j.1600-0587.2013.00127.x
Matias, M.G., Gravel, D., Guilhaumon, F., Desjardins-Proulx, P., Loreau, M., Münkemüller, T., Mouquet, N., 2014. Estimates of species extinctions from species-area relationships strongly depend on ecological context. Ecography 37, 431–442, http://dx.doi.org/10.1111/j.1600-0587.2013.00448.x
Merow, C., Smith, M.J., Edwards, T.C., Guisan, A., McMahon, S.M., Normand, S., et al., 2014. What do we gain from simplicity versus complexity in species distribution models? Ecography 37, 1–15, http://dx.doi.org/10.1111/ecog.00845
Ministerio de Medio Ambiente, 2006. Tercer Inventario Forestal Nacional, Barcelona. Dirección General para la Biodiversidad, Madrid.
Moudrý, V., Šímová, P., 2012. Influence of positional accuracy, sample size and scale on modelling species distributions: a review. Int. J. Geogr. Inform. Sci. 26, 2083–2095, http://dx.doi.org/10.1080/13658816.2012.721553
Pagel, J., Schurr, F.M., 2012. Forecasting species ranges by statistical estimation of ecological niches and spatial population dynamics. Glob. Ecol. Biogeogr. 21 (2), 293–304, http://dx.doi.org/10.1111/j.1466-8238.2011.00663.x
Papadimitriou, C.H., 1994. Computational Complexity. John Wiley and Sons Ltd.
Pearson, R.G., Dawson, T.P., 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Glob. Ecol. Biogeogr. 12, 361–371, http://dx.doi.org/10.1046/j.1466-822X.2003.00042.x
Pearson, R.G., Thuiller, W., Araújo, M.B., Martinez-Meyer, E., Brotons, L., McClean, C., Miles, L., Segurado, P., Dawson, T.P., Lees, D.C., 2006. Model-based uncertainty in species range prediction. J. Biogeogr. 33, 1704–1711, http://dx.doi.org/10.1111/j.1365-2699.2006.01460.x
R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.
Randin, C.F., Dirnböck, T., Dullinger, S., Zimmermann, N.E., Zappa, M., Guisan, A., 2006. Are niche-based species distribution models transferable in space? J. Biogeogr. 33 (10), 1689–1703, http://dx.doi.org/10.1111/j.1365-2699.2006.01466.x
Segurado, P., Araújo, M.B., 2004. An evaluation of methods for modelling species distributions. J. Biogeogr. 31, 1555–1568, http://dx.doi.org/10.1111/j.1365-2699.2004.01076.x
Smith, A.B., Santos, M.J., Koo, M.S., Rowe, K.M.C., Rowe, K.C., Patton, J.L., et al., 2013. Evaluation of species distribution models by resampling of sites surveyed a century ago by Joseph Grinnell. Ecography 36, 1017–1031, http://dx.doi.org/10.1111/j.1600-0587.2013.00107.x
Swets, J.A., 1988. Measuring the accuracy of diagnostic systems. Science 240 (4857), 1285–1293, http://dx.doi.org/10.1126/science.3287615

Please cite this article in press as: García-Callejas, D., Araújo, M.B., The effects of model and data complexity on predictions from species distributions models. Ecol. Model. (2015), http://dx.doi.org/10.1016/j.ecolmodel.2015.06.002

Thuiller, W., 2004. Patterns and uncertainties of species' range shifts under climate change. Glob. Change Biol. 10, 2020–2027, http://dx.doi.org/10.1111/j.1365-2486.2004.00859.x

Valladares, F., Matesanz, S., Guilhaumon, F., Araújo, M.B., Balaguer, L., Benito-Garzón, M., et al., 2014. The effects of phenotypic plasticity and local adaptation on forecasts of species range shifts under climate change. Ecol. Lett. 17, 1351–1364, http://dx.doi.org/10.1111/ele.12348

Vapnik, V., 1998. Statistical Learning Theory. Wiley.

Wilson, D.R., Martinez, T.R., 1997. Improved heterogeneous distance functions. J. Artif. Intel. Res. 6, 1–34.

Zanini, F., Pellet, J., Schmidt, B.R., 2009. The transferability of distribution models across regions: an amphibian case study. Divers. Distrib. 25, 469–480, http://dx.doi.org/10.1111/j.1472-4642.2008.00556.x.

Please cite this article in press as: García-Callejas, D., Araújo, M.B., The effects of model and data complexity on predictions from species distributions models. Ecol. Model. (2015), http://dx.doi.org/10.1016/j.ecolmodel.2015.06.002