A strict maximum likelihood explanation of MaxEnt, and some implications for distribution modelling

Rune Halvorsen

Rune Halvorsen, Department of Research and Collections, Natural History Museum, University of Oslo, P.O. Box 1172 Blindern, NO-0318 Norway.

sommerfeltia

Halvorsen, R. 2012. A maximum likelihood explanation of MaxEnt, and some implications for distribution modelling. – Sommerfeltia 36: 1-132. Oslo. ISBN 82-7420-050-0 . ISSN 0800-6865. DOI: 10.2478/v10208-011-0016-2.

Distribution modelling – research with the purpose of modelling the distribution of observable objects of a specific type – has become established as an independent branch of ecological science, with strong proliferation of approaches and methods in recent years. Since it was first made available to distribution modellers in 2004, the maximum entropy modelling method (MaxEnt) has established itself as a state-of-the-art method for distribution modelling. Default options and settings in the user-friendly Maxent software has become established as a standard practice for distribution modelling by MaxEnt.

A mini-review of 87 recent publications in which MaxEnt was used with empirical data to model distributions showed that the 'standard MaxEnt practice' is followed by a large majority of users and questioned by few. However, the review also provides indications that MaxEnt models obtained by the standard practice are sometimes overfitted to the data used to parameterise the model; examples of cases in which simpler MaxEnt models with predictive performance do exist. Results of the review motivate strongly for a better understanding of the ecological implications of the maximum entropy principle, as a basis for choosing MaxEnt options and settings.

This paper provides a thorough explanation of MaxEnt for ecologists, ending with a set of suggestions for improvements to the current practice of distribution modelling by MaxEnt. The explanation for MaxEnt given in the paper differs from previous explanations by being based on the maximum likelihood principle and by being based upon a gradient analytic perspective on distribution modelling. Four new findings are particularly emphasised: (1) that a strict maximum likelihood explanation of MaxEnt is possible, which places MaxEnt among regression methods in the widest sense; (2) that the true degrees of freedom for the residuals of a MaxEnt null model is $N - n$, the difference between the number of background and the number of presence observations used in the modelling; (3) that likelihood-ratio and F-ratio tests can be used to compare nested MaxEnt models; and (4) that subset selection methods are likely to be preferential to shrinkage methods for model selection in MaxEnt. Methods for internal model performance assessment, model comparison, and interpretation of MaxEnt model predictions (MaxEnt output), are described and discussed. Two simulated data sets are used to explore and

VERSITA

illustrate important issues relating to MaxEnt methodology.

Arguments for development of a generally applicable ‘consensus MaxEnt practice’ for spatial prediction modelling are given, and elements of such a practice discussed. Five main additions or amendments to the ‘standard MaxEnt practice’ are suggested: (1) flexible, interactive tools to assist deriving of variables from raw explanatory variables; (2) interactive tools to allow the user freely to combine model selection methods, methods and approaches for internal model performance assessment, and model improvement criteria, into a data-driven modelling procedure, (3) integration of independent presence/absence data into the modelling process, for external model performance assessment, for model calibration, and for model evaluation; (4) new output formats, notably a probability-ratio output format which directly expresses the ‘relative suitability of one place vs. another’ for the modelled target; and (5) development of options for discriminative use of MaxEnt, i.e., use of with presence/absence data. The most important research needs are considered to be: (1) comparative studies of strategies for construction of parsimonious sets of derived variables for use in MaxEnt modelling; and (2) comparative tests on independent presence/absence data of the predictive performance of MaxEnt models obtained with different model selection strategies, different approaches for internal model performance assessment, and different model improvement criteria.

Keywords: Distribution modelling; *F*-ratio test; Gradient analysis; MaxEnt; Maximum likehood; Model calibration; Model evaluation; Model selection; Regularisation

Abbreviations: AIC = Akaike’s information criterion; AUC = area under the (ROC) curve; BIC = Bayesian information criterion; BRT = boosted regression trees; C = set of binary variables derived from one categorical explanatory variable; D = deviation type of derived variables; DM = distribution modelling; DV = derived (explanatory) variable; DVMT = derived variable main type; DVT = derived variable type; ERM = ecological response modelling; EV = explanatory variable; FP = frequency of presence; FPR = false positive rate; FTVA = fraction of total variation accounted for; GAM = generalised additive models; GLM = generalised linear models; H = hinge type of derived variables; HF = forward hinge subtype of derived variables; HOF = Huisman-Olff-Fresco (models); HR = reverse hinge subtype of derived variables; ISDV = individually significant derived variable; K-S test = Kolmogorov-Smirnov test; L = linear derived variable; LM = linear regression model; M = monotonous type of derived variables; MARS = multivariate adaptive regression splines; MaxEnt = maximum entropy (model); Maxent = maximum entropy modelling software (Phillips et al. 2006, Phillips & Dudík 2008, Phillips 2011); O = covariance type of derived variables; OC = optimisation criterion; OP = observed presence vector; OPA = observed presence or absence vector; P = product derived variable; P/A = presence/absence; PCA = principal component analysis; PE = prediction error; PL = penalised likelihood; PO = presence-only; PPM = projective distribution modelling; PPP = predicted probability of presence; Q = quadratic derived variable; rDV = ‘raw’ derived variable; ROC = receiver operating characteristic (curve); RPPP = relative predicted probability of presence; SE = standard error (of the mean); SPM = spatial prediction modelling; T = threshold derived variable; TPR = true positive rate; V = variance derived variable; VA = variation accounted for; VC = variable contribution (to model); X = complex spline transformation type of derived variable.

Contents

INTRODUCTION

SETTING THE SCENE: OVERVIEW OF THE DISTRIBUTION MODELLING PROCESS

Distribution modelling (DM) comprises ‘research with the purpose of modelling the distribution of observable objects of a specific type’ [Halvorsen (2012), modified from Elith et al. (2006)]. DM has proliferated strongly in recent years, with respect to the diversity of available approaches [see Franklin (2009) and Peterson et al. (2011)] and the rate by which new papers are published (Lobo et al. 2010). Distribution modelling has deep roots in ecology and biogeography, as shown by the central position of the gradient analytic perspective in the theoretical foundation of DM (Austin 2007, Halvorsen 2012).

The distribution modelling process can be described as a 12-step process (Halvorsen 2012), as illustrated in Fig. 1 [see Halvorsen (2012) for definitions of terms and for further explanation of each step]:

1. Problem formulation and specification
2. Collection of raw data for the modelled target. The modelled target is often a species, but DM methods equally well apply to other natural phenomena, as exemplified by

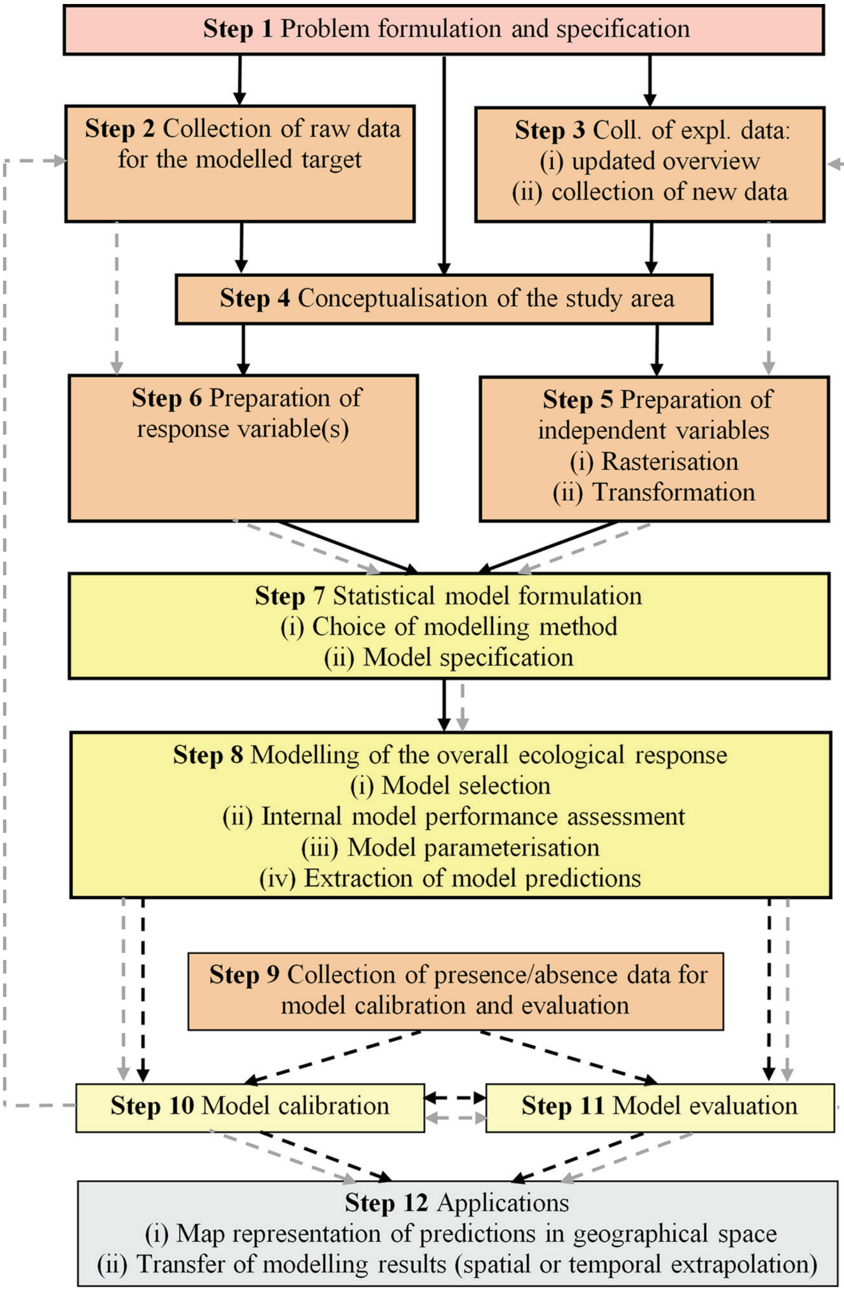

Step 1 Problem formulation and specification
Step 2 Collection of raw data for the modelled target
Step 3 Coll. of expl. data: (i) updated overview (ii) collection of new data
Step 4 Conceptualisation of the study area
Step 6 Preparation of response variable(s)
Step 5 Preparation of independent variables (i) Rasterisation (ii) Transformation
Step 7 Statistical model formulation (i) Choice of modelling method (ii) Model specification
Step 8 Modelling of the overall ecological response (i) Model selection (ii) Internal model performance assessment (iii) Model parameterisation (iv) Extraction of model predictions
Step 9 Collection of presence/absence data for model calibration and evaluation
Step 10 Model calibration
Step 11 Model evaluation
Step 12 Applications (i) Map representation of predictions in geographical space (ii) Transfer of modelling results (spatial or temporal extrapolation)

Fig. 1. Overview of the distribution modelling process, emphasising interdependencies between the 12 analytic steps. Steps are grouped into three composite steps, 'ecological model' (red background), 'data model' (orange background), and 'statistical model' (yellow background), in accordance with Austin (2002). Steps that are mandatory for a study to be distribution modelling, are indicated by thick borders. Steps involved in re-iteration of the model are indicated by gray lines. Broken lines indicate optional pathways. From Halvorsen (2012: Fig. 8).

species richness (Wohlgemuth et al. 2008, Aranda & Lobo 2011, Dubuis et al. 2011), nature types (Dobrowski et al. 2008, Danz et al. 2011), landforms (Hjort & Marmion 2009), and ecological processes such as wildfire (Parisien & Moritz 2009) and abandonment of agricultural practices (Gellrich & Zimmermann 2007). The term 'modelled target' is used for the studied entity throughout this paper.

3. Collection of explanatory data
4. Conceptualisation of the study area (as a rasterised geographical space)
5. Preparation of derived variables from 'raw' explanatory variables, by (i) rasterisation, followed by (ii) transformation into derived variables (also termed 'derived predictor variables') of which one or more may be derived from each explanatory variable
6. Preparation of response variable(s) from raw data for the modelled target
7. Statistical model formulation, by (i) choice of modelling method and (ii) model specification
8. Modelling of the overall ecological response of the modelled target, i.e., (i) model selection; (ii) internal model performance assessment; (iii) model parameterisation; and (iv) extraction of model predictions
9. Collection of presence/absence data for model calibration and evaluation.
10. Model calibration, a term used here for the process by which the numerical accuracy of model predictions is assessed: 'the level of agreement between predictions generated by a model and actual observations' (Pearce & Ferrier 2000b). 'Re-calibration' of models with PO data, as discussed by Phillips & Elith (2010), differs fundamentally from model calibration by use of P/A data. Typically, calibration implies that relative predicted probabilities of presence (RPPP) values obtained by use of presence-only (PO) data for the response in DM are brought onto a probability scale. Calibration is performed *a posteriori*, i.e., after modelling of the overall ecological response in Step 8, by calibration modelling.
11. Model evaluation, i.e., assessment of model performance by use of data not *directly* used to parameterise the model (Guisan & Zimmermann 2000).
12. Applications.

The 12 steps can be grouped into three composite steps in accordance with Austin (2002: 101) as follows (see Fig. 1): Step 1 belongs to 'ecological model', i.e., 'theory to be used or tested'; Steps 2–6, and 9, belong to 'data model', i.e., 'collection and measurement of ... data'; and Steps 7, 8, 10, and 11 belong to 'statistical model', i.e., 'the statistical theory and methods used'. Most of Steps 2–12 benefit strongly from being informed by basic ecological theory.

Steps 1–8 are essential for a study to belong to distribution modelling (DM) as defined above, i.e., as a study in which the primary response variable describes a distribution (Steps 2, 4 and 6), with explanatory variables that represent environmental gradients and are recorded or estimated for all grid cells within the extent of the study (Steps 3–5), and in which the modelled property is the overall ecological response (performance in environmental variables space; Steps 7–8). *Step 8, modelling of the overall ecological response, places DM unambiguously among gradient analysis techniques* as defined by ter Braak & Prentice (1988).

The outcome of the distribution modelling process most strongly hinges on the statistical model chosen by the modeller (Step 7) and his or her choice of options and settings for the modelling process (Step 8), although also other steps, such as data collection (Steps 2–3) and data preparation (Steps 5–6) are important (Halvorsen 2012). Modelling of modelled target's overall ecological response (Steps 7–8) is challenging for several reasons of which the most important is likely to be the variability of response-curve shapes – between modelled targets, for each modelled target between different environmental complex-gradients, and for each modelled target and complex-gradient between geographical areas and over time [see Halvorsen (2012) and references quoted therein]. Furthermore, the performance of modelling methods, options and settings interact with idiosyncratic properties of the modelled target in the study area to determine the outcome of DM (Elith et al. 2006, Guisan et al. 2007, Tognelli et al. 2009, Bedia et al. 2011).

The modelling purpose dictates what is regarded as a good distribution model and, hence, determines which model performance criteria are appropriate (Step 8,ii and Step 11). Halvorsen (2012) distinguishes between three main purposes of distribution modelling:

1. Ecological response modelling (ERM), distribution modelling with the main purpose of modelling the relationship between the performance of a modelled target and a set of explanatory variables, to find and understand general patterns in the modelled target's overall ecological response to the supplied explanatory variables. ERM thus addresses relationships in environmental variables (or ecological) conceptual spaces (Halvorsen 2012). ERM purposes can be divided into two sub-categories:
    a. Specific-purpose ecological response modelling, i.e., to describe and understand distributional variation at relevant scales, with regard to a specific set of explanatory variables.
    b. General-purpose ecological response modelling, i.e., to describe and understand distributional variation at relevant scales, without regard to a specific set of explanatory variables.
2. Spatial prediction modelling (SPM), distribution modelling with the main purpose of optimising the fit between model predictions and the true distribution of the modelled target's performance in the study area in the time interval data were collected
3. Projective distribution modelling (PPM), distribution modelling with the main purpose to transfer model predictions to a spatiotemporal setting different from the one at which the data used for modelling were collected. PPM purposes comprise variation from pure spatial-transfer distribution modelling, by which model predictions are to be projected into an area different from the area in which data were collected (the study area) but with environmental variation within the range spanned by the study area, and pure temporal-transfer distribution modelling, by which model predictions are to be projected into the study area, respectively, to 'new-context distribution modelling', by which projections are to be made into an environmental (e.g., climatic) scenario different from the range of environmental variation of the study area.

While SPM models are benchmarked by their capability for accurate prediction of independent presence/absence (P/A) evaluation data from the study area (Austin 2007, Lahoz-Monfort et al. 2007, Raes & ter Steege 2007, Veloz 2009, Edrén et al. 2010, Edvardsen et al. 2011, Halvorsen 2012), ERM cannot be evaluated by performance on data and have to be judged by ecological realism (Austin 2007), i.e., by their ability to summarise generalisable relationships between the modelled target and the environment, transferable in space and time (Halvorsen 2012). Independent P/A data can be used to evaluate spatial-transfer PPM models while empirical data for evaluation of temporal-transfer PPM models typically cannot be obtained. Evaluability by predictive performance on empirical data is an important difference between SPM and ERM (Araújo & Guisan 2006, Jiménez-Valverde et al. 2008, Braunisch & Suchant 2010, Warren & Seifert 2011, Halvorsen 2012).

Modelling of the overall ecological response (Step 7–8 in the six-step DM process) is a special case of statistical modelling. Statistical modelling can be defined as the process of finding the most parsimonious model (Hastie et al. 2009), i.e., the model which best combines simplicity in terms of number of model parameters with high predictive power (SPM) and/or expression of generally valid relationships between the performance of the modelled target and the environment (ERM). During the search for the most parsimonious model, the modeller makes many important decisions. Examples of such decisions are: which to choose among the large number of statistical modelling methods available for DM; which to choose among the numerous options and settings for the chosen method [see Franklin (2009) for an overview]; which methods to use for model comparison and evaluation; how to choose explanatory variables; and how to transform these in the most appropriate way [e.g., see Steyerberg et al. (2000), Burnham & Anderson (2002), Reineking & Schröder (2006), Zuur et al. (2007), Hastie et al. (2009), and Halvorsen (2012) for overview]. While there is growing consensus about which methods generally give the best SPM models (e.g., Elith et al. 2006, Mateo et al. 2010, Rebelo & Jones 2010, Rupprecht et al. 2011), choosing strategy for transformation of explanatory variables (Step 5,ii) and model selection (Step 8,i) have remained controversial issues for which clear guidelines still do not exist (Guisan & Zimmermann 2000, Pearce & Ferrier 2000a, Araújo & Guisan 2006, Anderson & Gonzalez 2011, Merckx et al. 2011). The importance of model selection is emphasised by Warren & Seifert (2011) who state that '... models that are inappropriately complex or inappropriately simple show reduced ability to infer habitat quality, reduced ability to infer the relative importance of variables in constraining the species performance-environment distributions, and reduced transferability to other time periods.'

MAXENT MODELLING OF SPECIES DISTRIBUTIONS

Distribution modelling branched off from mainstream gradient analysis in the 1990s (Guisan & Zimmermann 2000), developed into a more or less independent branch of ecological science in the 2000s (Franklin 2009, Halvorsen 2012), and now makes up the core of the new research field of conservation biogeography (Whittaker et al. 2005, Franklin 2010). Halvorsen (2012) argues that this new branch of ecological science still lacks 'the firm foothold offered by a strong theoretical foundation: in-depth understanding of the major processes and mechanisms that are responsible for observed patterns, built upon a conceptual basis that consists of precisely defined terms'. In support for this claim he cites the disagreement among distribution modellers on the relevance of ecological niche theory for DM, the lack of consensus on performance of several modelling methods and their options and on model selection and evaluation procedures, and the tendency for development of 'schools' with different research paradigms, a characteristic typical of research areas tenuously rooted in theory (Austin 2007). However, since 2004 a strong trend in DM has been the steadily increasing use of maximum entropy (MaxEnt) modelling (Elith et al. 2011) for Step 7,i in the 12-step DM process. MaxEnt's growing popularity is a result of easy access to user-friendly software and consistently high performance of the MaxEnt method in comparative tests of SPM methods.

MaxEnt was first proposed as a method for distribution modelling in 2004 (Phillips et al. 2004). The method has been freely available to users from day one via the Maxent software [note that 'MaxEnt' is used throughout this paper to denote the statistical method while 'Maxent' is used for the software] which has frequently been updated with new options (Phillips et al. 2006, Phillips & Dudík 2008, Elith et al. 2010, Phillips 2011). Recently, initiatives have been taken to integrate Maxent software with the R programming environment (Warren et al. 2010, Hijmans & Elith 2011, Phillips 2011).

Shortly after MaxEnt was first introduced to distribution modellers, the method was ranked top three in the most comprehensive test of DM methods published to date (Elith et al.

2006). The other top-ranked methods were boosted regression trees (BRT; De´ath 2007, Elith et al. 2008) and multivariate adaptive regression splines (MARS; Friedman 1991, Leathwick et al. 2006). Elith et al. (2006) used presence/absence evaluation data to test the predictive performance of 15 methods or variants of methods for 226 species of plants and animals. Later comparative studies of DM methods have confirmed the results of Elith et al. (2006), ranking MaxEnt as the best method or among the best (Hernandez et al. 2006, Guisan et al. 2007, Sérgio et al. 2007, Wang et al. 2007, Wisz et al. 2008, Roura-Pascual et al. 2009, Tognelli et al. 2009, Václávik & Meentemeyer 2009, Veloz 2009, Williams et al. 2009, Mateo et al. 2010, Rebelo & Jones 2010, Rupprecht et al. 2011). Inferior performance of MaxEnt has only been found in a few, exceptional cases [Peterson et al. (2007), Rota et al. (2011); but see Phillips (2008)].

Maximum entropy modelling is not one single method but rather a family of methods which originated in statistical mechanics more than 50 years ago (Jaynes 1957a, 1957b). Since then MaxEnt has undergone considerable development (e.g., Jaynes 2003), including adaptation to different research questions in several branches of science (see Phillips et al. 2004, Dudík & Phillips 2007). In ecology, MaxEnt is used among others for testing community assembly rules (e.g., Shipley et al. 2006, Roxburgh & Mokany 2010, Shipley 2010).

MaxEnt was introduced to distribution modellers as a machine-learning approach, i.e., as a method based on 'the idea ... to estimate a target probability distribution by finding the probability distribution of maximum entropy (i.e., that is most spread out, or closest to uniform), subject to a set of constraints that represent our incomplete information about the target distribution' (Phillips et al. 2004, 2006, Dudík et al. 2007). Phillips & Dudík (2008) characterise MaxEnt as 'robust Bayes estimation ... explained from a decision theoretic perspective', and Dudík & Phillips (2009) affiliate MaxEnt with 'robust Bayesian decision theory'.

In most publications intended for distribution modellers MaxEnt is characterised as a method for analysis of presence-only (PO) or presence/background data (Phillips et al. 2006, Phillips & Dudík 2008, Elith et al. 2011). MaxEnt modelling of presence/background data is the generative approach to MaxEnt modelling, in machine learning language also known as one-class estimation (Dudík & Phillips 2009). With presence-only data MaxEnt provides estimates of the probability that *one specific* presence cell, selected at random from all presence cells, is grid cell *i* (Phillips et al. 2006). The maximum entropy principle does, however, also apply to presence/absence (P/A) data. This is the discriminative approach to MaxEnt modelling or two-class estimation (Berger et al. 1996). With presence/absence data MaxEnt provides estimates of a quantity that is monotonously related to the probability of presence of the modelled target in grid cell *i*, conditioned on the environmental conditions (Dudík & Phillips 2009).

CURRENT PRACTICE: A MINI-REVIEW OF DISTRIBUTION MODELLING STUDIES USING MAX-ENT

MaxEnt is a flexible modelling method with many options and settings that may be specified, or 'tuned', by the user (see Phillips 2011). Many of these options and settings can be tuned independently of each other, and innumerable MaxEnt models can therefore be constructed for the same data set. Detailed technical explanations of each of these options and settings from a machine-learning perspective are available, as exemplified by the description of the algorithm used in the Maxent software and the proof for its convergence to a unique solution provided by Dudík et al. (2007). However, machine-learning theory and concepts are outside the experience of most ecologists, and the theoretical framework of machine learning cannot easily be transferred to the ecological realities dealt with in distribution modelling – modelled target's responses to the environment and distributions in geographical and environmental variables spaces (Elith et al. 2011). Accordingly, choosing the optimal combination of MaxEnt options and settings when the distribution of a specific modelled target is to be modelled for a specific purpose, often by use of a data set given *a priori*, and understanding how these choices interact with the ecological interpretability of modelling results, have remained major and to a large extent unresolved challenges for practical distribution modellers (e.g., VanDerWal et al. 2009a, Anderson & Gonzalez 2011, Warren & Seifert 2011).

As a preamble to this study of distribution modelling by MaxEnt, I performed a mini-review of 87 distribution modelling studies published in international journals 2006–11, in which Maxent software was applied to empirical data (Table 1). Nearly one half of the reviewed studies (41 %) neither included explanations of how the derived variables used in parameterisation of the MaxEnt model were derived from the raw explanatory variables (Step 5,ii of the 12-step DM process) nor reported settings for the model selection procedure (Step 8,i), and 33 % of the studies (only partly overlapping with the former group) failed to contain an explicit statement of which version of Maxent software was used. Most likely all of these studies used Maxent default values for all options and settings, including the so-called ℓ_1-regularisation method by which model selection (Step 8,i), internal model performance assessment (Step 8,ii) and model parameterisation (Step 8,iii) are combined into one complex procedure.

Before models are built, Maxent software transforms all continuous explanatory variables [‘environmental layers’ in the terminology used by Phillips (2011)] into derived variables of (up to) four different types [termed ‘features’ by Phillips & Dudík (2008)]: in addition to the raw explanatory variable (the ‘linear’ variable type) the ‘quadratic’, the ‘hinge’ and the ‘threshold’ types. The two last-mentioned types are exceptional in that a very high number of variables of each type can be derived from each explanatory variable (Dudík et al. 2007). Maxent model complexity is regulated in the first place by applying (default) threshold minimum values for the number of presence observations required for derived variables of the ‘quadratic’, ‘hinge’ and ‘threshold’ types to enter the model. Threshold minimum numbers are set separately for each type. Default values for these thresholds were overruled (mostly by replacement with more restrictive settings) only in 16 % of the studies considered for the mini-review, and the default value for constructing interaction variables (‘product features’) was overruled in 13 % of the studies only (the overlap with the former group was high).

A pre-selection of explanatory variables before default options for ‘feature’ construction and regularisation were applied, was found in 25% of the studies. Methods used for variable pre-selection include inspection of correlation patterns and removal of variables that are strongly correlated with other variables (Gibson et al. 2007, Young et al. 2009, Gaikwad et al. 2011, Ko et al. 2011, Marino et al. 2011, Parisien & Moritz 2011, Rupprecht et al. 2011), and variable reduction by principal component analysis (PCA) ordination (Tognelli et al. 2009; Verbruggen et al. 2009, Williams et al. 2011) or generalised linear modelling (GLM; Wollan et al. 2008, Bedia et al. 2011). In a few, exceptional cases derived variables were manually selected also for the final model by forward, backward, or forward-backward selection, using Maxent variable diagnostics to compare competing models (Parolo et al. 2008, Lahoz-Monfort et al. 2007, Bradley et al. 2010, Merckx et al. 2011).

The default MaxEnt procedure for preventing models to be overfitted to the training data, ℓ_1-regularisation (the lasso; Tibshirani 1996), is one among several methods for model shrinkage, which in turn is one among several strategies for model selection (a full explanation of model selection in MaxEnt is given in the ‘Theory’ chapter). The *modus operandi* of ℓ_1-regularisation is to reduce (shrink) the absolute values of model parameters (Phillips & Dudík 2008, Hastie et al. 2009) in a process that is very flexible and requires tuning of the regularisation parameter(s) which determine the degree to which parameters are penalised for being large (Reikeking

& Schröder 2006). The default Maxent procedure is to determine regularisation parameters separately for each derived variable from the number of presence observations in the data set, the type of variable ('feature type'), and its variance (Phillips & Dudík 2008). Strong faith in ℓ_1-regularisation with default tuning as a guarantor against overfitting is expressed in many studies (e.g., Hernandez et al. 2006, Parolo et al. 2008, Wisz et al. 2008, Wollan et al. 2008, Merckx et al. 2011, Rupprecht et al. 2011). Accordingly, default regularisation settings are reported to be overruled by the user only in five of the studies (6 %) included in the mini-review, of which stronger than default regularisation was reported in three (Lamb et al. 2008, Elith et al. 2010, Naimi et al. 2011). None of the authors of studies included in the mini-review reported not to have used ℓ_1-regularisation.

The mini-review clearly shows that use of default options and settings in Maxent software constitutes a well-established standard practice for distribution modelling by MaxEnt. I will use the term 'standard Maxent practice' for distribution modelling by MaxEnt which makes use of following options (see Elith et al. 2011, Phillips 2011):

1. Automatic transformation of raw explanatory variables to derived variables of specific types ('feature types' in Maxent terms), governed by number of presence observations
2. Interactions automatically included, by product variables ('product features'), when the number of presence observations exceeds a default threshold
3. Model selection by ℓ_1-regularisation (the lasso) to prevent overfitting

This standard MaxEnt practice is followed by a large majority of users of Maxent software and has only occasionally been questioned [but see Anderson & Gonzalez (2011) and Warren & Seifert (2011)].

A CRITIQUE OF THE 'STANDARD MAXENT PRACTICE' FOR DISTRIBUTION MODELLING

The mini-review reveals that most MaxEnt users apply default settings of the Maxent software. However, based on indications that models obtained by standard MaxEnt practice may be overfitted to the data, some authors have recently questioned uncritical use of these default options and settings. In a comparative study of methods for model comparison and regularisation parameters Warren & Seifert (2011) conclude that 'at present little guidance is available for setting the appropriate level of regularisation, and the effects of inappropriately complex or simple models are largely unknown'. In their detailed study of the rare species *Cryptotis meridensis* Anderson & Gonzalez (2011) report that individual tuning of regularisation parameters enhances the models' predictive ability compared to default settings. Furthermore, Raes & ter Steege (2007) and Merckx et al. (2011) report tendencies for Maxent models with default settings, trained on 80 or more presence observations (the minimum number of presence observations at which 'threshold' and 'product' variable types are by default allowed to enter MaxEnt models), to be overfitted to the training data. Merckx et al. (2011) suggest that default values for the regularisation multiplier and threshold minimum values for transformation of explanatory variables into the derived variables that are used to parameterise the model should be reconsidered or, simply, that more strict rules for pre-selection of variables should be applied.

In DM contexts, the concept of overfitting can only be precisely defined by taking modelling purpose into account (Halvorsen 2012): Halvorsen (2012) defines an overfitted model as 'a distribution model that fits more complex overall response curves than appropriate, given the modelling purpose'. Furthermore, Halvorsen (2012) recognises three types of overfitting:

1. ***Type I overfitting***, i.e., that a more complex model has lower predictive performance on independent data than a simpler model.
2. ***Type II overfitting***, i.e., that a more complex model is similar (in the meaning 'not significantly better') in predictive performance on independent data than a simpler model.
3. ***Type III overfitting***, i.e., that a more complex model with higher predictive performance on independent data than a simpler model fails to fit realistic overall ecological response curves.

Type-I and Type-II overfitted models are inferior regardless of modelling purpose, the Type-II overfitted model because the simpler model is better according to the principle of parsimony, while Type-III overfitting is relevant to ERM models only.

These three definitions of 'overfitting types' all require information on the 'predictive performance on independent data', i.e., results of model evaluation (Step 11 in the 12-step DM process). Overfitting of Types I and II of MaxEnt models for the SPM purpose can be revealed in several ways, of which one of the most frequently used are comparison of the area under the receiver operating (ROC) curve (AUC; Hanley & McNeil 1982, Pearce & Ferrier 2000b) between simpler and more complex models (Merckx et al. 2011): much lower AUC values for complex models on evaluation data than on training data (i.e., the data used to parameterise the model), and reversed ranking of models by AUC from higher AUC of more complex than of simpler models on training data to the converse on evaluation data, are indications that the more complex models are overfitted. Thus, Parolo et al. (2008) found that an eight-variable MaxEnt model obtained for 60 presence observations of the plant species *Arnica montana* by use of default Maxent settings had lower AUC value (AUC = 0.864) on set-aside evaluation data than a manually pruned three-variable model (AUC = 0.888) while the two models were ranked differently by AUC on training data (AUC = 0.941 and 0.924, respectively). Svenning et al. (2008) concluded from visual inspection of map representations of model predictions in geographical space that a MaxEnt model with 12 explanatory variables was overfitted and inferior to a model with three manually selected variables, despite the former model had slightly higher AUC than the latter on training data (AUC = 0.767 and 0.751, respectively). Wollan et al. (2008) found in preliminary analyses with 75 explanatory variables that MaxEnt models with default Maxent settings tended to be strongly overfitted, and used logistic regression for pre-selection of variables before MaxEnt modelling. Finally, Yost et al. (2008) found a very small difference in AUC between models evaluated by repeated resubstitution of data between a full model with seven variables and a two-variable model obtained by sequential backward elimination of variables included in the full model.

Two results obtained by applying standard Maxent practice to empirical data have, in particular, triggered my suspicion that default Maxent options and settings result in overfitted models more often than previously anticipated: (1) The tendency for overall response curves produced by Maxent to be very complex (e.g., Dudík et al. 2007: Fig. 8; Elith et al. 2011: Fig. 5, Tab. S5-2), with shapes that deviate strongly from the smooth, symmetric or skewed, unimodal or truncated unimodal responses to major complex-gradients expected from gradient analytic theory (see Halvorsen 2012 and references cited therein). (2) The large number of derived variables with nonzero parameters that are typically listed in the *NN.lambdas* (*NN* is the name of the modelled target) output file from Maxent software. This file contains a list of all derived variables created during the modelling process and their parameters; all variables with nonzero parameters are included in the model. One example is the study by I. Auestad et al. (unpubl.

results) in which seven continuous explanatory variables were represented in the model by 110 derived variables with nonzero parameters. Visual inspection of this final model revealed clear signs that the model was overfitted: very strong local variation in the predicted response and very complex overall response curves to environmental gradients.

These concerns raise several questions: How complex are the published MaxEnt models? How many derived variables are included in the final models? How large are the model parameters, and how many hinge- and threshold-type variables are used to represent each explanatory variable? The mini-review (Table 1) failed to reveal one single study in which the number of derived variables or the values of parameters in the final MaxEnt model was reported; the methodological study by Warren & Seifert (2011) was not included in the mini-review. The practice in distribution modelling by MaxEnt not to report properties of the final model other than evaluation results and map representations of predictions strongly contrasts the practice in ecological modelling by standard statistical tools like GLM for other purposes, by which analysis and documentation of the model itself is regarded as essential (cf. Zuur et al. 2007, Hastie et al. 2009).

The mini-review (Table 1) suggests that a better understanding of effects of choice of options and settings in MaxEnt modelling is needed and, in particular, that the standard Maxent practice should be carefully evaluated. If the predictive power of MaxEnt models may, at least in some cases, be further improved by making simpler models, an important side effect will be that models optimised for spatial prediction (SPM) may also express more generally valid relationships between the modelled target and the environment, as required of ERM and PPM models. These simpler MaxEnt models may, in case, have the additional benefit that they open for the dual purpose of combining good spatial predictions with improved understanding of the ecological determinants of observed distributions (Halvorsen 2012).

Table 1. Mini-review of current practice in distribution modelling by MaxEnt, using the Maxent software: characteristics of 87 recent publications in international journals (marked by asterisk in the reference list).

Characteristic	% of studies
Reference to version of Maxent software lacking	33
Explanation of method used to derive derived variables from explanatory variables *and* regularisation settings lacking	41
Departure from automatic procedure for transformation of single explanatory variables to derived variables	16
Default threshold for including derived variables of the product type (≥ 80 presence observations) overruled by the user	13
Default regularisation settings overruled by the user	6
ℓ_1-regularisation not used	0
Pre-selection of explanatory variables prior to Maxent modelling in which default options are used for formation of derived variables and regularisation	25
Explicit statement of the number of derived variables (parameters) in the final MaxEnt model	1
Explicit specification of the final MaxEnt model, including model parameters	0

THE NEED FOR A MAXIMUM LIKELIHOOD EXPLANATION OF MAXENT

The prospect that the current practice of MaxEnt modelling might be improved triggered my curiosity for what really goes on in the 'MaxEnt black box', and, notably, how to understand the method and its options and settings from an ecological point of view. The machine-learning perspective, which is adopted in most explanations of MaxEnt (e.g., Phillips et al. 2006, Phillips & Dudík 2008), is generally recognised as difficult to translate into ecological terms (Elith et al. 2011). However, Phillips et al. (2006) mentioned that 'Maxent has strong similarities to some existing methods for modelling species distributions, in particular, generalised linear models (GLMs) [and] generalised additive models (GAMs)', but also wrote that MaxEnt 'is not as mature a statistical method as GLM or GAM ... [and that] there are fewer guidelines for its use in general ... and fewer methods for estimating the amount of error in a prediction'. In recent years several authors have explored the similarity between MaxEnt and regression methods (e.g., Gibson et al. 2007, Willems & Hill 2011). Elith et al. (2011) were the first to explain MaxEnt as a statistical method for modelling the overall ecological response in (continuous) environmental variables space: using a combination of regression (GLM and GAM) modelling terminology and Bayesian statistical concepts they characterised MaxEnt as a method that 'minimizes the relative entropy between two probability densities (one estimated from the presence data and one, from the landscape) defined in covariate space'.

Phillips et al. (2006) and Dudík et al. (2007) point to the existence of an element of maximum likelihood estimation implicit in the MaxEnt method, and thus open for the possibility that the maximum likelihood perspective can be used to understand MaxEnt. While the machine-learning perspective emphasises MaxEnt's ability to provide robust estimates of relative 'habitat' suitabilities in abstract geographical space [see Halvorsen (2012) for explanation of conceptual spaces], statistical principles like maximum likelihood estimation emphasise the method's ability to model relative suitabilities in environmental variables space (Elith et al. 2011). Elith et al. (2011) argue that 'for many users, this [statistical] viewpoint is likely to be a more accessible way to understand the [MaxEnt] model than previous ones that rely on machine learning concepts'.

Proper understanding of complex methods requires in-depth knowledge of the principles on which the methods are based and how the methods are linked with basic theory (Økland 1990, 2007, Austin 1999, 2007). Halvorsen (2012) argues that the appropriate theoretical background for distribution modelling is the gradient analytic perspective (Halvorsen 2012). Maximum likelihood estimation stands out as a promising methodological platform for linking MaxEnt to ecological theory.

FOCUS AND AIMS

The main aims of this paper are: (1) to provide a strict maximum likelihood explanation of the MaxEnt method, including options and settings implemented in the Maxent software (Phillips et al. 2004, 2006, Phillips & Dudík 2008, Phillips 2011) and other integrated tools (e.g., Franklin 2009, Halvorsen 2012), and to link this explanation with the gradient analytic perspective on distribution modelling; (2) to provide a description of the MaxEnt method, including options and settings, that is sufficiently detailed to allow the interested reader to follow all steps; (3) to illustrate MaxEnt by simple worked examples, (4) to address still unsettled issues in MaxEnt modelling; (5) to discuss practical implications of the strict maximum likelihood explanation of MaxEnt; and, (6) to discuss additions to, or changes of, the standard Maxent practice for distribution modelling.

This paper is organised in three parts: (1) a theory chapter that provides a strict maximum likelihood explanation of MaxEnt, its options and settings, and that links MaxEnt with the gradient analytic perspective on distribution modelling; (2) a worked examples chapter in which two simple simulated data sets are used to illustrate the method (with options, settings and tools) and to address important, still unsettled issues in MaxEnt methodology; and (3) a final chapter in which the use of MaxEnt in distribution modelling is discussed with the aim of suggesting additions to, or changes of, this practice, and to point to issues in need of further research.

THEORY

This chapter provides a full description of the MaxEnt method for DM, based on maximum likelihood principles. The chapter is structured according to the 12-step process of the DM (Fig. 1), starting with the data model (Steps 2–6) which is followed by the statistical model (Steps 7, 8, 10 and 11). A full account of the notation used in this chapter is given in Appendix 1.

DATA MODEL

Distribution modelling takes place within the bounds of a geographically delimited study area. For the purpose of distribution modelling by MaxEnt we assume that the study area is conceptualised as a rasterised geographical space (Step 4 of the 12-step DM procedure), i.e., that the area is divided into N_T equal-sized, quadratic grid cells or pixels in which the modelled target and explanatory variables are recorded. The set of N_T discrete observation units is denoted $\boldsymbol{D}_T = \{d_1, ..., d_i, ..., d_n, ..., d_N, ..., d_{N_T}\}$. The grid-cell edge length defines the unit grain, or observation unit, size of the study. If N_T is very large (typically > 10 000; Phillips & Dudík 2008), computation time can be reduced without loss of model precision by using a subset of the N_T observations for modelling. The theory is equally applicable to modelling situations in which a subset $\boldsymbol{D}$ of $\boldsymbol{D}_T$, with N observations, is used. Therefore, if not otherwise is stated, it will be assumed that $\boldsymbol{D} = \boldsymbol{D}_T$ and that $N = N_\mathrm{T}$, i.e., that modelling makes use of all grid cells in the study area.

The 'raw' explanatory data set used for DM (Steps 3 and 5,i), which is denoted $\mathbf{Z}$, consists of an $N \times s$ matrix with elements z_{ij} that are values for the jth explanatory variable in grid cell d_i, $d_i \in \boldsymbol{D}$. The matrix $\mathbf{Z}$ has the s explanatory variables $\mathbf{Z}_j$ as columns; $\mathbf{Z} = [\mathbf{Z}_1, ..., \mathbf{Z}_S]$, $\mathbf{Z}_{\boldsymbol{j}} = [z_{1j}, ..., z_{Nj}]^\mathrm{T}$. The row vectors of $\mathbf{Z}$, i.e., the values for the s explanatory variables recorded for grid cell i, are denoted Z_i.

Response variables for DM by MaxEnt (Step 2) can be of two principally different kinds: presence-only (PO) and presence/absence (P/A) data. MaxEnt models parameterised by use PO data are referred to as generative MaxEnt models while models parameterised by P/A data are referred to as discriminative MaxEnt models (Dudík & Phillips 2009). However, for most cases of practical distribution modelling only PO data are available (e.g., Franklin 2009, Robertson et al. 2010, Feeley & Silman 2011, Niamir et al. 2011) and generative MaxEnt is therefore the main focus of this paper.

The rasterised observed presence (OP) response vector $\boldsymbol{C} = [c_1, ..., c_i, ..., c_N]^T$ (Step 6) consists of records of observed presence in n grid cells ($c_i = 1$) while information about presence or absence is lacking for the remaining $N - n$ cells. The latter, which are referred to as *uninformed background cells*, are given the value $c_i = 0$. We adopt the sorting convention that grid cells $i = 1, ..., n$ are the n observed presence cells while cells $i = n + 1, ..., N$ are the $N - n$ uninformed background cells:

$$c_i = \begin{cases} 1 \text{ for } i = 1, \dots, n \\ 0 \text{ for } i = n + 1, \dots, N \end{cases} \quad (1)$$

The rasterised observed presence or absence (OPA) response vector will, when available, be referred to as $\boldsymbol{B} = [b_1, ..., b_i, ..., b_N]^T$. This vector consists of records of presence in n_+ presence grid cells ($b_i = 1$) and of absence in $N - n_+$ absence cells ($b_i = 0$). As above, we adopt the sorting convention that the first n_+ cells of $\boldsymbol{B}$ are the n_+ presence cells and. In cases where both PO and P/A response vectors exist, the first n of the n_+ presence cells are observed presence cells ($c_i = 1$). The vectors $\boldsymbol{B}$ and $\boldsymbol{C}$ differ by the identity of the $N - n$ grid cells in which presence is not recorded ($c_i = 0$): $c_i = 1 \Rightarrow b_i = 1$ (n grid cells) and $b_i = 0 \Rightarrow c_i = 0$ ($N - n_+$ grid cells), while $c_i = 0$ ($N - n$ grid cells) corresponds to an unknown value of b_i. I use bold-face italicised capital letters for sets and vectors with N_T (= N) elements, i.e., that contain values for all grid cells in the study area. Bold-face normal letters are used for matrices.

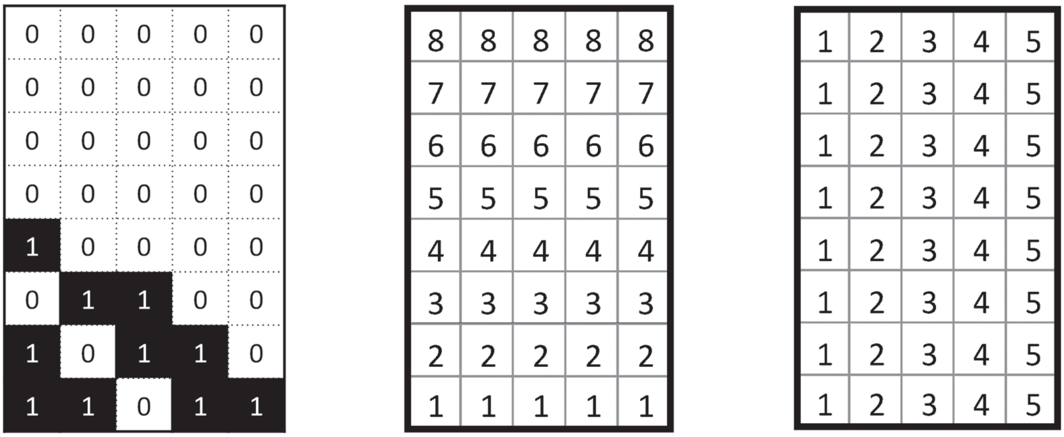
A small simulated example data set will be used to illustrate the data model and exemplify some aspects of MaxEnt theory. This data set, which will be denoted 1^*, is a subset of example data set 1 used in the 'Worked examples' chapter. The rasterised study area for data set 1^* consists of 40 grid cells and is denoted $\boldsymbol{D} = \{d_1, ..., d_i, ..., d_{40}\}$. The grid cells are arranged in 8 rows × 5 columns (Fig. 2a). A simulated target species 'Sp' is observed in $n = 10$ (25 %) of the total $N = 40$ grid cells in $\boldsymbol{D}$ (Fig. 2a). No information is available about eventual presence or absence of Sp in the remaining $N - n = 30$ uninformed background grid cells. The environmental data set $\mathbf{Z}$ consists of two explanatory variables ($s = 2$), of which both are recorded for each grid cell in $\boldsymbol{D}$; $\boldsymbol{Z}_1 = [z_{1,1}, ..., z_{1i}, z_{1,40}]^T$ and $\boldsymbol{Z}_2 = [z_{2,1}, ..., z_{2i}, z_{2,40}]^T$. $\boldsymbol{Z}_1$ indexes northing ('Y coordinate') in the rasterised geographical space representation of the study area (Fig. 2b) while $\boldsymbol{Z}_2$ indexes easting ('X coordinate') in this space (Fig. 2c). The PO response vector $\boldsymbol{C} = [c_1, ..., c_i, ..., c_{40}]^T$ for Sp has the value $c_i = 1$ in 10 cells which, by applying the sorting convention, are indexed from 1 to 10. The remaining 30 cells are uninformed background cells for which $c_i = 0$.

OUTLINE OF THE MAXENT STATISTICAL MODELLING PROCESS

Basically, the statistical model comprises Steps 7, 8, 10 and 11 in the 12-step DM process (Halvorsen 2012). However, in generative MaxEnt the response variable $\boldsymbol{C}$ is modelled not directly as a response to s explanatory variables $\mathbf{Z}_j$ (EVs) but to a set $\mathbf{X}$ of m derived variables $\boldsymbol{X}_k$ (DVs) obtained from $\boldsymbol{Z}_j$ by transformation. The term 'derived variable' (DV) is used here in exactly the same meaning as the term 'feature' in studies by Phillips et al. (2006), Dudík & Phillips (2007), Phillips & Dudík (2008), and Elith et al. (2011). The general relationship between DVs $\boldsymbol{X}_k$ and EVs $\boldsymbol{Z}_j$ is given by transformation and back-transformation functions h and h^{-1} defined as follows:

$$\boldsymbol{X}_k = h_k(\mathbf{Z}) \Leftrightarrow \boldsymbol{Z}_j = h_j^{-1}(\mathbf{X}) \quad (2)$$

In this theory chapter, elements of the statistical modelling process by MaxEnt are ordered by their sequence of appearance in the modelling process. Because the transformation from EVs



0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
1	0	0	0	0
0	1	1	0	0
1	0	1	1	0
1	1	0	1	1

8	8	8	8	8
7	7	7	7	7
6	6	6	6	6
5	5	5	5	5
4	4	4	4	4
3	3	3	3	3
2	2	2	2	2
1	1	1	1	1

1	2	3	4	5
1	2	3	4	5
1	2	3	4	5
1	2	3	4	5
1	2	3	4	5
1	2	3	4	5
1	2	3	4	5
1	2	3	4	5

Fig. 2. Example data set 1 (and 1*). (a) The study area ***D*** which is rasterised into 40 grid cells of which the modelled target is recorded in 10 (black cells). Values for the observed presence vector ***C*** in each cell is shown. (b) Observed values z_{1i} for explanatory variable (EV) $\boldsymbol{Z}_1$ in the 40 grid cells. (c) Observed values z_{2i} for EV $\boldsymbol{Z}_2$ in the 40 grid cells.

to DVs (Step 5,ii in the 12-step DM process) often, like in the Maxent software (Phillips et al. 2006), is carried out an integrated part of the statistical modelling process, the transformation step is described here first, followed by a detailed description of the MaxEnt statistical model. The following notation and terms are used for the DVs: $\mathbf{X} = [\boldsymbol{X}_1, ..., \boldsymbol{X}_k, ..., \boldsymbol{X}_m]$ denotes the matrix of values for *m* DVs in *N* observation units. The column and row vectors of **X** are denoted $\boldsymbol{X}_k = [x_{1k}, ..., x_{ik}, ..., x_{Nk}]^{\mathrm{T}}$ and $X_i = [x_{i1}, ..., x_{ik}, ..., x_{im}]$, respectively.

The description of the MaxEnt statistical model starts with a detailed description of the 'core' of the modelling process, which is followed by 'other important aspects of the MaxEnt statistical model'. Starting with the 'core' of the MaxEnt modelling process, a brief outline of DM by MaxEnt is given here to motivate the way issues are sorted on these two main groups as well as within each group in this paper.

Formulated in the most generally way, a MaxEnt distribution model *Q* describes the relationship between one response variable (RV), $\boldsymbol{Y} = [y_1, ..., y_i, ..., y_N]^{\mathrm{T}}$, and one or more EVs, $\boldsymbol{Z}_j$, that are represented by DVs, $\boldsymbol{X}_k$, by mathematical functions of the exponential family in which the exponent is linear in $\boldsymbol{X}_k$. Choice of MaxEnt as modelling method (Step 7,i) therefore also specifies the model (Step 7,ii).

A DM ideal for applied purposes should provide predictions of the probability of presence (PPP) of the modelled target, i.e., the *real* probability that the modelled target is present in grid cell *i*; $y_i = \Pr(b_i = 1 | \mathbf{Z})$ (Edwards et al. 2005, Guisan et al. 2006, Edvardsen et al. 2011, Halvorsen 2012). However, MaxEnt model estimates $\boldsymbol{Q} = [q_1, ..., q_i, ..., q_N]^{\mathrm{T}}$ can be interpreted as estimates of PPP if and only if the prevalence of the modelled target in the study area, i.e., the modelled target's frequency of presence (Hirzel et al. 2006, Halvorsen 2012), is known. This condition can only be directly met by discriminative MaxEnt, i.e., MaxEnt models obtained by use of the observed presence or absence (OPA) vector ***B*** as RV (Ward et al. 2009). In contrast, 'raw' estimates, or predictions, from generative MaxEnt models are *relative* predicted probabilities of presence (RPPP) for the *N* observation units used for model parameterisation (Phillips et al.

2006, Phillips & Dudík 2008, Ward et al. 2009). The 'raw' RPPP values, the q_i's, of generative MaxEnt models Q, are by definition *probabilities* that *one specific* presence cell i_0, selected at random from all n_+ true presence cells, happens to be grid cell i:

$$q_i = \Pr(i = i_0 \mid b_{i_0} = 1) \tag{3}$$

This definition gives the vector $\boldsymbol{Q}$ of generative MaxEnt model estimates for the N grid cells the property of a discrete probability distribution, i.e., that $\sum_{i=1}^{N} q_i = 1$. The q_i's are functions of $m + 1$ mode parameters $\Theta = [\theta_0, \theta_1, ..., \theta_k, ..., \theta_m]^{\mathrm{T}}$ and the EVs $\boldsymbol{Z}_j$ as represented by the m DVs, conditioned on $\sum_{i}^{N} q_i = 1$:

$$q_i = g_\Theta{}' \,(Z_i) = g_\Theta \,(X_i) \tag{4}$$

where $g_\Theta{}'$ is the MaxEnt model expressed as a function of the 'raw' EVs and the parameter vector Θ, and g_Θ is the model expressed as a function of the DVs and Θ. Because the DVs and not the EVs themselves are used in generative MaxEnt modelling, model estimates q_i are expectations of the relationship between the response variable $\boldsymbol{C}$ and $\boldsymbol{X}_k$. Relationships between $\boldsymbol{C}$ and $\mathbf{Z}_j$ are obtained by inserting (2) in (3):

$$q_i = \Pr\,(i = i_0 \mid b_{i_0} = 1; \Theta, \mathbf{X}) = \Pr\,(i = i_0 \mid b_{i_0} = 1; \Theta, h_k\,(\mathbf{Z})) \tag{5}$$

Because the property modelled by generative MaxEnt models is not the response variable $\boldsymbol{Y}$ itself, but the derived property of the response variable given by expression (3), the notation $\boldsymbol{\Pi} = [\pi_1, ..., \pi_i, ..., \pi_N]^{\mathrm{T}}$ is used instead of $\boldsymbol{Y}$ for the true discrete probability distribution estimated by $\boldsymbol{Q}$. Each π_i expresses the probability that a randomly selected presence grid cell i_0: $\varphi_{i0} = 1$, is cell i.

The predictions q_i from a MaxEnt model Q, expressed in terms of X_i by (4), can be back-transformed to a function of Z_i by use of (2):

$$q_i = g_\Theta{}' \,(Z_i) = g_\Theta{}' \,(h_k^{-1}\,(X_i)) \tag{6}$$

This 'core' of the modelling process comprises the model specification step (Step 7,ii) and Steps (8,i–iii) of the modelling of the overall ecological response (Step 8). These steps are addressed *en suite* in this paper because in MaxEnt pre-selected methods for internal model performance assessment (Step 8,ii) are used in a tightly integrated process of model selection (Step 8,i) and model parameterisation (Step 8,iii).

'Other important aspects of the MaxEnt statistical model' comprises three steps that may be carried out after a model Q has been obtained, i.e., *a posteriori*, with the purpose of enhancing the practical value of modelling results:

1. *Interpretation and transformation of model predictions* (Step 8,iv). The 'raw' MaxEnt estimates can be transformed into several 'output formats' (Phillips et al. 2006) which fit different modelling purposes, partly because they address relationships in different conceptual spaces. Vectors $\boldsymbol{C}$, $\mathbf{Z}_j$, and $\boldsymbol{Q}$ all contain values for N discrete points in geographical space [see Halvorsen (2012) for definitions of conceptual spaces]. However, the geographical space in which MaxEnt operates is an abstract geographical space because the georeference of observation units is not explicitly used in the modelling. Because MaxEnt makes use of the EV data matrix $\mathbf{Z}$, modelling can with equal right be considered as taking place in a discrete environmental variables space, i.e., with $\mathbf{Z}_j$

as axes. MaxEnt models can also be used for estimation or prediction in continuous environmental variables space, i.e., for a combination of EVs without reference to a specific raster.

2. *Model calibration* (Step 10). Many practical purposes require 'upgrading' of generative MaxEnt modelling results from RPPP estimates to PPP estimates by an *a posteriori* model calibration step accomplished by use of independently collected P/A data (Step 9).
3. *Model evaluation* (Step 11). The tendency for PO-data to be burdened with strong sampling bias (Araújo & Guisan 2006, Hortal et al. 2008, Loiselle et al. 2008, Robertson et al. 2010, Wolmarans et al. 2010, McCarthy et al. 2011) makes model evaluation by use of a set of P/A data collected independently of the data used to parameterise the model an essential step in the DM process (Araújo et al. 2005, Austin 2007, Raes & ter Steege 2007, Veloz 2009, Edvardsen et al. 2011, Halvorsen 2012).

Only the first two of these steps, which make use of procedures specific to each DM method, are specifically dealt with in this paper.

TRANSFORMATION OF EXPLANATORY VARIABLES INTO DERIVED VARIABLES (STEP 5,iii)

One or more derived variables (DVs) $\boldsymbol{X}_k$ can be derived from explanatory variables (EVs) $\boldsymbol{Z}_j$ by transformation. There are no inherent restrictions in the MaxEnt method with respect to categories of transformation functions *h* that can be used to derive DVs from EVs: in principle, all kinds of continuous and discontinuous transformations ('smoothers') available for GLM (Phillips et al. 2006) and GAM (e.g., see Wood 2006) can be used for MaxEnt.

Five types of transformation functions are available in the Maxent software for continuous EVs (Phillips & Dudík 2008, Elith et al. 2011). These can be sorted into three main types: two types of *continuous variables*; two types of discontinuous *spline variables*; and one type of *interaction variables* which combines two or more EVs into one DV. Categorical EVs make up a main category on its own. This set of six main types of transformation functions is, however, not exhaustive. Additional variable types of obvious relevance to DM are therefore proposed in this paper, giving a total of nine types of variables in four main types (see Table 2):

1. Continuous derived variables
   a. The *linear* (L) type, i.e., the untransformed EV itself
   b. The *monotonous* (M) type, i.e., any strictly monotonous transformation of the continuous variable $\boldsymbol{Z}_j$. The *quadratic* (Q) variable of Phillips & Dudík (2008), i.e., $\boldsymbol{Z}_j$ squared, is one example of an M-type DV. By including M-type variables, nonlinear relationships between the modelled target's response and major complex-gradients are explicitly taken into account in the DM process.
   c. The *deviation* (D) type, expressing the deviation of $\boldsymbol{Z}_j$ from the mean value of $\boldsymbol{Z}_j$ over the *m* observed presence sites, denoted $\bar{z}_j^*$. D-type variables can be expressed on the general form $(z_{ij} - \bar{z}_j^*)^a$ where *a* is a scalar [expression (7) in Table 2]. The 'variance' (V) variable, which was proposed for use in DM by Gastón & García-Viñas (2011), corresponds to $a = 2$. The V variable is analogous with a variance because it is based upon squared deviations. By including D variables, the tolerance of a modelled target with unimodal ecological response to a major complex-gradient is explicitly taken into account in the DM process.

Table 2. Transformation of explanatory variables (EVs) into derived variables (DV): characteristics of derived variable (DV) main types (DVMTs) and types (DVTs) relevant for MaxEnt modelling. The transformation procedure occurs in two steps, of which only the first step, transformation into ‘raw’ derived variables (rDVs) $\boldsymbol{X}_k'$, is shown in the rightmost table column. The DVs $\boldsymbol{X}_k$ are obtained by linear ranging of rDVs onto a [0,1] scale. * = DVTs not currently implemented in Maxent software.

DVMT	DVT		Description	Interpretation	Transformation function for rDVs	
	Code	Term				
conti-nuous	L	linear	the continuous EV $\boldsymbol{Z}_j$ itself	models the response to the EV itself	$x_{ik}' = h_L(z_{ij}) = z_{ij}$	
conti-nuous	M	mono-tonous	a monotonous, continuous transformation of the continuous EV $\boldsymbol{Z}_j$	models the response to a nonlinear transformation of the EV; the quadratic (Q) variable obtained as the square of $\boldsymbol{Z}_j$ is a special case	$x_{ik} = h_M(z_{ij}) = \mathrm{f}\,(z_{ij})$ where f is a continuous function	
conti-nuous	D*	deviation	the continuous EV $\boldsymbol{Z}_j$, centred on the mean for observed presence grid cells, raised to the power a	takes the tolerance of the species with respect to the EV explicitly into account by modelling the response to the spread of z_{ij} around the mean value for observed presence grid cells,; the V (variance) variable, which is obtained for a = 2, is a special case	$x_{ik}' = h_D(z_{ij}) = (z_{ij} - \bar{z}_j^*)^a$	(7)
spline	HF	forward hinge	a continuous EV $\boldsymbol{Z}_j$ transformed to a linear spline of order two	models the response to a piecewise linear spline with one knot (the point z_{0j}) above which $\boldsymbol{X}_k$ is a linear function of $\boldsymbol{Z}_j$ and below which $\boldsymbol{X}_k$ is set equal to 0	$x_{ik}' = h_{HF}(z_{ij}) = \begin{cases} 0 \text{ if } z_{ij} < z_{0j} \\ \dfrac{z_{ij} - z_{0j}}{\max(z_{ij}) - z_{0j}} \text{ if } z_{ij} \geq z_{0j} \end{cases}$	(8)
spline	HR	reverse hinge	a continuous EV $\boldsymbol{Z}_j$ transformed to a linear spline of order two	models the response to a piecewise linear spline with one knot (the point z_{0j}) below which $\boldsymbol{X}_k$ is a linear function of $\boldsymbol{Z}_j$ and above which $\boldsymbol{X}_k$ is set equal to 0	$x_{ik}' = h_{HR}(z_{ij}) = \begin{cases} \dfrac{z_{0j} - z_{ij}}{z_{0j} - \min(z_{ij})} \text{ if } z_{ij} \leq z_{0j} \\ 0 \text{ if } z_{ij} > z_{0j} \end{cases}$	(9)

Table 2 (Continued).

DVMT	DVT		Description	Interpretation	Transformation function for rDVs
	Code	Term			
spline	T	threshold	binary transformation of a continuous EV $\boldsymbol{Z}_j$	piecewise constant spline with one knot (discontinuity point z_{0j}) below which $\boldsymbol{X}_k$ is set equal to 0 and above which $\boldsymbol{X}_k = 1$; models the proportion (frequency) of presence grid cells with $z_{ij} \geq z_{0j}$	$x_{ik}{}' = h_T(z_{ij})$ $= \begin{cases} 1 \text{ if } z_{ij} \geq z_{0j} \\ 0 \text{ if } z_{ij} < z_{0j} \end{cases}$ (10)
spline	X*	complex spline	a continuous EV $\boldsymbol{Z}_j$ transformed to a linear spline of order three or higher	models a continuous or discontinuous, complex, response to an EV	$x_{ik} = h_M(z_{ij}) = \mathrm{f}\,(z_{ij})$ where f is a discontinuous spline function of order three or higher
inter-action	P	product	the product of two continuous EVs $\boldsymbol{Z}_j$ and $\boldsymbol{Z}_v$	models the response to the product of two continuous DVs	$x_{ik}{}' = h_P(z_{ij}, z_{iv}) = z_{ij} \cdot z_{iv}$ (11)
inter-action	O*	covariance	the product of two continuous EVs $\boldsymbol{Z}_j$ and $\boldsymbol{Z}_v$, centered on the respective means for observed presence grid cells	takes the interaction (covariance) between the modelled target's tolerances to two EVs explicitly into account	$x_{ik}{}' = h_O(z_{ij}, z_{iv})$ $= (z_{ij} - \bar{z}_j^*) \cdot (z_{iv} - \bar{z}_v^*)$ (12)
binary set	C	binary	m_j binary DVs, one for each factor level u of a categorical EV $\boldsymbol{Z}_j$; each DV expresses if factor level u is recorded in cell i or not	models the proportion (frequency) of presence grid cells for each factor level u	m_j binary DVs, one for each factor level u: $x_{ik} = h_C\,(z_{ij})$ $= \begin{cases} 1 \text{ if } z_{ij} = u \\ 0 \text{ if } z_{ij} \neq u \end{cases}$ (13)

2. Spline variables, which make use of knots, values for the EV below and above which values are transformed by different functions (e.g., Zuur et al. 2007).
    a. The *hinge* (H) type, i.e., a piecewise linear spline transformation of order two, i.e., with one knot. The knot separates a portion of the EV with linear response from a portion with x_{ik} = 0, i.e., from which the modelled target is expected to be absent. Two subtypes of H-type DVs can be recognised: 'forward hinge' (HF) with x_{ik} = 0 for $z_{ij} > z_{0j}$, the value of the knot [expression (8) in Table 2], and 'reverse hinge' (HR) with x_{ik} = 0 for $z_{ij} < z_{0j}$ [expression (9) in Table 2]. Hinge-type DVs account for the situation where the modelled target reaches a tolerance limit with respect to a major complex-gradient within the range of variation encountered in the study area.
    b. The *threshold* (T) type, i.e., a piecewise linear spline transformation of order two by which a knot separates two portions of the EV with constant response (presence, x_{ik} = 1, in one and absence, x_{ik} = 0, in the other). T-type DVs [expression (10) in Table 2] account for situations with threshold response (Halvorsen (2012), i.e., abrupt changes in the modelled target's overall ecological response to a major complex-gradient.
    c. The *complex spline transformation* (X) type, i.e., transformations into piecewise linear functions of order three or higher. X-type variables open for modelling complex discontinuous overall ecological responses to a major complex-gradient.
3. Interaction variables
    a. The *product* (P) type, i.e., the product of two EVs, or, equivalently, two L-type DVs [expression (11) in Table 2].
    b. The *covariance* (O) type, which is the parallel to the V variable of the D type, is defined as the product of two continuous EVs $\boldsymbol{Z}_j$ and $\boldsymbol{Z}_v$, centred on the respective means for observed presence sites [expression (12) in Table 2]. This DV resembles a covariance by its multiplication of deviations from a mean. By including an O variable, interactions between responses to two complex-gradients is explicitly taken into account in the DM process. In principle, the O type can been defined more generally, opening for more complex relationships between two or more EVs in line with interactions of higher order than two.
4. Sets of binary variables (C), one set derived from each categorical EV. An EV with *u* factor levels is represented by *u* binary DVs [expression (13) in Table 2]

These types of transformations of continuous EVs are not clearly separated. For instance, the linear variable (L) can be considered as a special case of a monotonous (M) variable or as a forward hinge (HF) variable in which the knot is placed at the lower extreme (Elith et al. 2011).

While only one L-type, one Q and one V variable can be derived from each EV, many DVs of the H, T, M and D types that can be derived from each EV. The number of unique threshold-type variables that can be obtained from a continuous EV is bound above by N – 1 (Dudík et al. 2007), while there is no *a priori* upper bound on the number of hinge-type variables that can be derived. By including in the model many DVs of the threshold (T) and/or hinge (H) types from each of the continuous explanatory variables, and by moving the position z_{j0} of the knot, response curves of almost all shapes and complexities can be modelled.

The transformation procedure by which DVs are obtained from EVs consists of two steps:

1. *Transformation into 'raw' derived variables* (rDVs), $\boldsymbol{X}_k{}' = [x_{1k}{}', ..., x_{ik}{}', ..., x_{Nk}{}']^{\mathrm{T}}$, as outlined above and operationalised by transformation formulae given in the rightmost column of Table 2.
2. *Ranging into derived variables* (DVs), by linear rescaling (*ranging*; Økland 1990) of each rDV into a new variable with values in the range [0,1], by

$$x_{ik} = \frac{(x_{ik}{}' - \mathrm{x}_{k,\min}{}')}{(x_{k,\max}{}' - \mathrm{x}_{k,\min}{}')} \tag{14}$$

where x_{ik} denotes the DV and $x_{ik}{}'$ denotes the rDV, and $x_{k,\min}{}'$ and $x_{k,\max}{}'$ denote the minimum and maximum values of the latter.

rDVs of the HR, HF, T and C types are transformed directly into ranged DVs by the transformation functions given in Table 2. Ranging makes all DVs comparable by bringing them onto the same scale. By Maxent software, all DVs are ranged of is performed as an integrated part of Step 8 in the DM process.

Examples of DVs derived from the two EVs in example 1^* are shown in Table 3.

MODEL SPECIFICATION AND MODELLING OF THE OVERALL ECOLOGICAL RESPONSE (STEPS 7,ii AND 8,i–iii)

A maximum likelihood explanation of MaxEnt for distribution modelling

According to the maximum likelihood principle, the set parameter vector Θ of a generative model Q_{Θ} that maximises the likelihood of obtaining the vector $\boldsymbol{\Pi}$ of true, underlying values π_i of the modelled target is the best among all possible models (Hastie et al. 2009). Accordingly, a maximum likelihood solution to modelling of the overall ecological response implies finding the set of parameters Θ that maximises Pr_{Θ} ($\boldsymbol{\Pi}$ | $\mathbf{Z}$); the probability of $\boldsymbol{\Pi}$ given the environmental conditions $\mathbf{Z}$. Note that most statistical analyses rest on the assumption that observations of the response variable are independent and identically distributed, drawn from the population

Table 3. Values for derived variables (DVs) of different types, derived from explanatory variables (EVs) $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ in example data set 1^* (see Table 2 for explanation of types of DVs). Note that the DVs are ranged to a [0, 1] scale.

DV type	Knot	Value for EV $\boldsymbol{Z}_1$								Value for EV $\boldsymbol{Z}_2$				
		1	2	3	4	5	6	7	8	1	2	3	4	5
L	–	0.000	0.143	0.286	0.429	0.571	0.714	0.857	1.000	0.000	0.250	0.500	0.750	1.000
Q	–	0.000	0.020	0.082	0.184	0.327	0.510	0.735	1.000	0.000	0.063	0.250	0.563	1.000
V	–	0.028	0.000	0.028	0.111	0.250	0.444	0.694	1.000	0.429	0.036	0.000	0.321	1.000
HF	0.500	0.000	0.000	0.000	0.000	0.143	0.429	0.714	1.000	0.000	0.000	0.000	0.500	1.000
HR	0.500	1.000	1.000	1.000	1.000	0.857	0.571	0.286	0.000	1.000	1.000	1.000	0.500	0.000
T	0.500	0.000	0.000	0.000	0.000	1.000	1.000	1.000	1.000	0.000	0.000	1.000	1.000	1.000

of all possible observations. In the context of distribution modelling by MaxEnt, this implies that (1), the probability that one particular presence cell i_0, selected at random from all presence cells, is a specific grid cell i_a, is independent of (2), the probability that the cell i_0 is i_b, for all pairs of grid cells i_a and i_b.

When PO data are used for distribution modelling, we do not know if the modelled target is *really* present or absent (i.e., whether $b_{\mathrm{i}} = 1$ or $b_{\mathrm{i}} = 0$) in uninformed background ($c_{\mathrm{i}} = 0$) cells but, under the assumption that misidentifications and other errors are not present in the data, we can assume that the modelled target is present ($b_{\mathrm{i}} = 1$) in the n presence cells ($c_{\mathrm{i}} = 1$). The machine-learning explanation of MaxEnt emphasises that only reliable information shall be used to estimate $\boldsymbol{\Pi}$. The reliable information about sites where the modelled target is present consists of grid cells with observed presence ($c_{\mathrm{i}} = 1$); $c_{\mathrm{i}} = 1 \Leftrightarrow b_{\mathrm{i}} = 1$. Thus, from the machine-learning perspective, best fit to reliable PO data is obtained by the model $\mathrm{Q_s}$ which approximates $\boldsymbol{\Pi} = [\pi_1, ..., \pi_i, ..., \pi_N]^{\mathrm{T}}$ with the discrete probability distribution vector $\boldsymbol{Q}_{\mathrm{s}} = [q_1, ..., q_i, ..., q_N]^{\mathrm{T}}$, the elements of which are:

$$q_{\mathrm{S},i} = \begin{cases} \frac{1}{n} \text{ for } i = 1, ..., n \\ 0 \text{ for } i > n \end{cases} . \tag{15}$$

The vector $\boldsymbol{Q}_{\mathrm{s}}$ is an important reference distribution for MaxEnt modelling. From a machine-learning perspective, $\mathrm{Q_s}$ is the model with best fit to reliable information about the modelled target because it separates the n presence cells from the $N - n$ uninformed background cells. From a maximum likelihood estimation perspective, the model $\mathrm{Q_s}$ is the *saturated model*, i.e., the model that accounts for all variation in the response variable, conditioned on the uninformed background grid cells being treated as pseudo-absence observations, i.e., as surrogates for real absence observations. Another important reference model is the *null model Q_0*, the model for which the available explanatory variables are of no use, or are not used, for predicting the presence observations. The null model attributes to all grid cells the same probability of being a randomly selected presence grid cell i_0. With N grid cells, the null model is the model with elements $q_{0,i} = \frac{1}{N}$ for all i.

A basic principle in statistical modelling is to seek for the most parsimonious model, i.e., the model which best combines simplicity (in terms of m, the number of model parameters) and predictive ability (Hastie et al. 2009). The saturated model is usually very complex in terms of numbers of model parameters. Models that fit the data closely tend not to express general relationships between response and derived variables; they are *overfitted* and poorly suited for prediction (Halvorsen 2012). In our example data set 1* neither of explanatory variables $\boldsymbol{Z}_1$ or $\boldsymbol{Z}_2$ predict observed presences perfectly; perfect prediction is only possible by specifying the 10 combinations of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ that correspond to observed presence ($c_{\mathrm{i}} = 1$) while absence is specified for the 30 uninformed background grid cells.

With PO data, we do not know which among uninformed background cells i, $i > n$, that are presences and which are absences. However, we do know with absolute certainty that the modelled target is present in observed presence cells, i.e., cells in the set $\boldsymbol{D}_+$ ($c_{\mathrm{i}} = 1 \Rightarrow b_{\mathrm{i}} = 1$). *In MaxEnt, restriction to use of reliable information is accomplished by maximising the likelihood of the presence cells only*, i.e., to use $\boldsymbol{\Pi} = \boldsymbol{\Pi}_{\mathrm{S}}$, with elements $\pi_{\mathrm{S},i} = q_{\mathrm{S},i}$ given by expression (4), as the reference with which all models are compared. MaxEnt shall thus maximise the likelihood for $\boldsymbol{Q}$, given that the true probability distribution is $\boldsymbol{\Pi}_{\mathrm{S}}$, and $\boldsymbol{Q}$ shall be a vector of maximum likelihood estimates:

$$\boldsymbol{Q} = \hat{\boldsymbol{\Pi}} \Leftrightarrow q_i = \hat{\pi}_i \tag{16}$$

For a set of independent observations, the likelihood of obtaining one particular vector $\boldsymbol{Q}_\Theta = [q_1, ..., q_i, ..., q_N]^{\mathrm{T}}$, i.e., one specific parameterisation of the model, is the product of the likelihoods for the N q_i values (in example 1*, $N = 40$). The definition of q_i gives

$$\Pr(q_i) = \Pr(i = i_0 \mid b_{i0} = 1, \mathbf{X}, \Theta\,) = q_i \tag{17}$$

Since the model estimates $\boldsymbol{Q}_\Theta$ by definition is a probability distribution, the likelihood L_Θ of $\boldsymbol{Q}_\Theta$ is therefore obtained as the product of likelihoods for each observation i. By using (3), we obtain:

$$\begin{aligned} L_\Theta &= \textstyle\prod_{i=1}^{N} q_i \\ &= \textstyle\prod_{i=1}^{N} \Pr(i = i_0 \mid b_{i0} = 1, \mathbf{X}, \Theta) \\ &= \textstyle\prod_{i=1}^{n} \Pr(i = i_0 \mid b_{i0} = 1, \mathbf{X}, \Theta) \cdot \prod_{i=n+1}^{N} \Pr(i = i_0 \mid b_{i0} = 1, \mathbf{X}, \Theta) \\ &= \textstyle\prod_{i=1}^{n} q_i \cdot \prod_{i=n+1}^{N} q_i \\ &= \mathrm{L}_{\Theta+} \cdot \mathrm{L}_{\Theta-} \end{aligned} \tag{19}$$

In accordance with the maximum likelihood principle we seek the model $\boldsymbol{Q}_\Theta$ that maximises the fit of the data to $\boldsymbol{\Pi}_{\mathrm{S}}$, i.e., the model that maximises

$$\mathrm{L}_\Theta = \textstyle\prod_{i=1}^{n} \Pr\,(\pi_i = \frac{1}{n} \mid \mathbf{X}, \Theta\,) \cdot \prod_{i=n+1}^{N} \Pr\,(\pi_i = 0 \mid \mathbf{X}, \Theta\,) \tag{20}$$

Expression (20) is obtained by inserting (15) in (18).

The fundamental principle of generative MaxEnt modelling (Jaynes 1957a, 1957b, Dudík et al. 2007, Dudík & Phillips 2009, Shipley 2010), to maximise $L_{\Theta+}$ instead of L_Θ, implies that the model $\boldsymbol{Q}_\Theta$ that maximises the likelihood of the n presence sites is searched for rather than the model that maximises the likelihood of all N sites. This is a fundamental difference between MaxEnt and, e.g., GLM (Shipley 2010). The model that optimises $L_{\Theta+}$ is, however, likely to be close to the model that optimises L_Θ because improving $L_{\Theta+}$ by increasing the likelihood q_i for a presence grid cell i ($c_{\mathrm{i}} = 1$; $i \le n$) towards $\frac{1}{n}$ necessarily implies lowering of q_i for at least one uninformed background cell and, hence, improvement of Pr ($\pi_i = 0 \mid \mathbf{X}, \Theta\,$) and the likelihood $L_{\Theta-}$ for uninformed background cells ($c_{\mathrm{i}} = 0$; $i \ge n$).

Because $a > b$ implies that ln $a >$ ln b, the model parameter vector Θ that maximises $L_{\Theta+}$ also maximises ln $L_{\Theta+}$, given by

$$\begin{aligned} \ln \mathrm{L}_{\Theta+} &= \textstyle\sum_{i=1}^{n} \ln \Pr\,(\pi_i = \frac{1}{n} \mid \mathbf{X}, \Theta\,) \\ &= \textstyle\sum_{i=1}^{n} \ln q_i \end{aligned} \tag{21}$$

This expression is the *Kullback-Leibler divergence* (Kullback 1959), an information theoretic measure of the extent to which two vectors $\Pi^* = [\pi_1, ..., \pi_i, ..., \pi_n]^{\mathrm{T}}$ and $Q^* = [q_1, ..., q_i, ..., q_n]^{\mathrm{T}} = \Pr\,(\pi_i = \frac{1}{n} \mid \mathbf{X}, \Theta\,)$differ (Phillips et al. 2004, Dudík et al. 2007, Shipley 2010, Elith et al. 2011).

Since all estimated probabilities q_i are < 1 by definition, ln $q_i < 0$ for all i and hence ln $L_{\Theta+} < 0$. Maximising (21) is therefore equivalent to minimising –ln $L_{\Theta+}$. By convention, the quantity minimised in Maxent software is not –ln $L_{\Theta+}$ but this quantity divided by n, which we denote ln L_Θ or, equivalently, ln L_t (for model t), depending on the context:

$$\ln L_\Theta = -\frac{1}{n} \cdot \ln L_{\Theta+} \tag{22}$$

By inserting (21) in (22), and using (4), we obtain:

$$\ln L_\Theta = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \; (-\ln q_i) \qquad (23)$$
$$= \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \; \ln\left(g_\Theta\,(X_i)\right)$$

It has been proved (Della Pietra et al. 1997) that the MaxEnt distribution [the distribution that minimises (23)] is a Gibbs distribution:

$$q_i = g_\Theta\,(X_i) = e^{\theta_0 + \Sigma_{k=1}^{m} \theta_k x_{ik}} \qquad (24)$$

where θ_0 is a ‘normalising constant’ that ensures that the set of N q_i values satisfies the condition of summing to 1. Accordingly, the MaxEnt distribution is defined by the set of parameters Θ which minimises the negative normalised log-likelihood (Dudík et al. 2007). Combining expressions (23) and (24) gives:

$$\ln L_\Theta = -\tfrac{1}{n} \textstyle\sum_{i=1}^{n} \; \ln\left(e^{\theta_0 + \Sigma_{k=1}^{m} \theta_k x_{ik}}\right). \qquad (25)$$

This quantity is often termed (the empirical) *log loss* (Phillips et al. 2006, Dudík et al. 2007, Phillips & Dudík 2008).

The demand on θ_0 that the N q_i values shall sum to unity makes θ_0 dependent on the other parameters θ_k as follows:

$$\textstyle\sum_{i=1}^{N} \; q_i = 1$$
$$\textstyle\sum_{i=1}^{N} \; e^{\theta_0 + \Sigma_{k=1}^{m} \theta_k x_{ik}} = 1$$
$$\textstyle\sum_{i=1}^{N} \; e^{\theta_0}\, e^{\Sigma_{k=1}^{m} \theta_k x_{ik}} = 1 \qquad (26)$$

Solving for (26) for e^{θ_0} gives

$$e^{\theta_0} \; = \; \frac{1}{\sum_{i=1}^{N} \; e^{\Sigma_{k=1}^{m} \theta_k x_{ik}}} \qquad (27)$$

from which is follows that the constant θ_0 is given by

$$\theta_0 = -\ln\left(\textstyle\sum_{i=1}^{N} \; e^{\Sigma_{k=1}^{m} \theta_k x_{ik}}\right). \qquad (28)$$

The expression for log loss given by (25) can be simplified as follows:

$$\ln L_\Theta = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} (-\ln e^{\theta_0}) - \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \ln \; e^{\Sigma_{k=1}^{m} \theta_k x_{ik}}$$
$$= \tfrac{1}{n} \cdot \mathrm{n} \cdot (-\,\theta_0) - \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \; \sum_{k=1}^{m} \theta_k \, x_{ik}$$
$$= \ln\left(\textstyle\sum_{i=1}^{N} \; e^{\Sigma_{k=1}^{m} \theta_k x_{ik}}\right) - \textstyle\sum_{k=1}^{m} \theta_k \cdot \left(\tfrac{1}{n} \textstyle\sum_{i=1}^{n} x_{ik}\right)$$
$$= \ln\,(\textstyle\sum_{i=1}^{N} \; e^{\Sigma_{k=1}^{m} \theta_k x_{ik}} \;\; - \textstyle\sum_{k=1}^{m} \theta_k \cdot \overline{x}_k^{\,*} \qquad (29)$$

where $\overline{x}_k^{\,*}$ is the mean of derived variable $\boldsymbol{X}_k$ in the n observed presence cells ($c_\mathrm{i} = 1$; $i \leq n$).

Parameters $\Theta = [\theta_0, \theta_1, ..., \theta_k, ..., \theta_\mathrm{m}]^\mathrm{T}$ of the best model are found as the solutions of the m equations:

$$\frac{\partial \ln L_\theta}{\partial \ln \theta_k} \; = 0;\ \mathrm{k} = 1, ..., \mathrm{m} \qquad (30)$$

Differentiating (18) with respect to θ_k gives:

$$\frac{\partial \ln L_\theta}{\partial \ln \theta_k} = 0$$

$$\frac{\partial \left[\ln \left(\sum_{i=1}^{N} e^{\Sigma_{k=1}^{m} \theta_k x_{ik}}\right) - \sum_{k=1}^{m} \theta_k \cdot \bar{x}_k^{\,*}\right]}{\partial \theta_k} = 0$$

$$\frac{\partial \ln \left(\sum_{i=1}^{N} e^{\Sigma_{k=1}^{m} \theta_k x_{ik}}\right)}{\partial \theta_k} - \frac{\partial \sum_{k=1}^{m} \theta_k \cdot \bar{x}_k^{\,*}}{\partial \theta_k} = 0$$

$$\frac{\dfrac{\partial \ln \left(\sum_{i=1}^{N} e^{\Sigma_{k=1}^{m} \theta_k x_{ik}}\right)}{\partial \theta_k}}{\sum_{i=1}^{N} e^{\Sigma_{k=1}^{m} \theta_k x_{ik}}} - \bar{x}_k^{\,*} = 0$$

$$\frac{\sum_{i=1}^{N} \dfrac{\partial \left(e^{\Sigma_{k=1}^{m} \theta_k x_{ik}}\right)}{\partial \theta_k}}{\sum_{i=1}^{N} e^{\Sigma_{k=1}^{m} \theta_k x_{ik}}} - \bar{x}_k^{\,*} = 0$$

$$\frac{\sum_{i=1}^{N} x_{ik} e^{\Sigma_{k=1}^{m} \theta_k x_{ik}}}{\sum_{i=1}^{N} e^{\Sigma_{k=1}^{m} \theta_k x_{ik}}} - \bar{x}_k^{\,*} = 0$$

$$\frac{\sum_{i=1}^{N} x_{ik} e^{\Sigma_{k=1}^{m} \theta_k x_{ik}}}{\sum_{i=1}^{N} e^{\Sigma_{k=1}^{m} \theta_k x_{ik}}} = \bar{x}_k^{\,*}$$

$$\frac{1}{\sum_{i=1}^{N} e^{\Sigma_{k=1}^{m} \theta_k x_{ik}}} \cdot \sum_{i=1}^{N} x_{ik} e^{\Sigma_{k=1}^{m} \theta_k x_{ik}} = \bar{x}_k^{\,*} \qquad (31)$$

Inserting (27) in (31) gives:

$$e^{\theta_0} \cdot \sum_{i=1}^{N} x_{ik} e^{\Sigma_{k=1}^{m} \theta_k x_{ik}} = \bar{x}_k^{\,*}$$
$$\sum_{i=1}^{N} x_{ik} e^{\theta_0} e^{\Sigma_{k=1}^{m} \theta_k x_{ik}} = \bar{x}_k^{\,*}$$
$$\sum_{i=1}^{N} x_{ik} e^{\theta_0 + \Sigma_{k=1}^{m} \theta_k x_{ik}} = \bar{x}_k^{\,*} \qquad (32)$$

Finally, inserting (24) in (32) gives

$$\sum_{i=1}^{N} x_{ik}\, qi = \bar{x}_k^{\,*} \qquad (33)$$

Expression (33) provides the set of conditions that have to be satisfied for a model Q_M to be the maximum likelihood MaxEnt model: for all derived variables $\boldsymbol{X}_k$, the weighted sum of derived variable values x_{ik} over all N grid cells using q_i as weights, which we denote $\tilde{x}_k$, shall equal the average of $\boldsymbol{X}_k$ over the n observed presence cells ($c_\mathrm{i} = 1$), $\bar{x}_k^{\,*}$.

$$\tilde{x}_k = \bar{x}_k^{\,*} \qquad (34)$$

Because the q_i sum to unity, $\tilde{x}_k$ is the *weighted average* of $\boldsymbol{X}_k$, using q_i as weights.

The condition that $\tilde{x}_k$ shall be equal to $\bar{x}_k^{\,*}$, i.e., the mean of $\boldsymbol{X}_k$ in observed presence cells, applies to all derived variables regardless of type. The 'mean of $\boldsymbol{X}_k$' has different meanings for different variable types (Table 2): with linear (L) variables it is simply the mean of the ranged explanatory variable over observed presence cells; with quadratic (Q) variables it is the mean of the squared variable over observed presence cells; with variance (V) variables it is the average mean squared deviation from the mean over observed presence cells; with hinge (H) variables it is the mean of $\boldsymbol{X}_k$ over observed presence cells in the subset of grid cells in which $\boldsymbol{X}_k$ is linearly related to $\boldsymbol{Z}_j$; and with threshold (T) and categorical (C) variables it is the mean of the binary variables $\boldsymbol{X}_k$ in observed presence cells, i.e., the proportion, or frequency, of presence cells for which $x_k = 1$. Mean values over the 10 observed presence grid cells, $\bar{x}_k^*$, for each the six variables derived from explanatory variable $\boldsymbol{Z}_1$ in example 1* are: $\bar{x}_{1,\mathrm{L}}^* = 0.1429$, $\bar{x}_{1,\mathrm{Q}}^* = 0{,}0408$, $\bar{x}_{1,\mathrm{V}}^* = 0{,}0278$, $\bar{x}_{1,\mathrm{HF.5}}^* = 0$, $\bar{x}_{1,\mathrm{HR.5}}^* = 1$ and $\bar{x}_{1,\mathrm{T.5}}^* = 0$ (compare Figs 2a and 2b with Table 3).

In the original implementation of MaxEnt for distribution modelling by Phillips et al. (2004, 2006), explained by adopting a machine-learning perspective, the target distribution $\boldsymbol{Q}_\mathrm{M}$ is defined as the probability distribution $\boldsymbol{Q}$ of maximum entropy, subject to the constraint that the mean of each derived variable $\boldsymbol{X}_k$ weighted by $q_{i,}$ i.e., $\tilde{x}_k$, equals the empirical mean $\bar{x}_k^*$ in the subset of presence cells. This corresponds exactly to the condition given by (34), which was derived from the maximum likelihood principle. Furthermore, the machine-learning explanation of MaxEnt (Phillips et al. 2004, 2006) uses the maximum entropy principle to define as best the model which satisfies the constraints given by (34) and at the same time maximises the relative entropy given by:

$$H_j = -\sum_{i=1}^{N} q_i \cdot \ln q_i \tag{35}$$

Dudík et al. (2007) show that finding the solution of maximum relative entropy under the constraints given in (34) is equivalent to minimising the log loss given by (29).

The standard method for internal assessment of the performance of a model Q_t (e.g., Hastie et al. 2009) is to compare Q_t with the standard reference models, the saturated model Q_S that explains all variation in the data and fits the data perfectly, and the null model Q_0 that explains no variation and does not at all make use of explanatory data. The saturated MaxEnt model is defined above as the model which satisfies the condition that $\boldsymbol{Q}^* = [q_1, ..., q_i, ..., q_n]^\mathrm{T}$ equals $\boldsymbol{\Pi}^* = [\pi_1, ..., \pi_i, ..., \pi_n]^\mathrm{T}$. In accordance with (15), this is the model which predicts a value of $q_i = \pi_i = \frac{1}{n}$ for all i = 1, ..., n and, since the q_i's sum to 1, a value of $q_i = \pi_i = 0$ for all $i = n + 1, ...,$ N. Accordingly, the log loss of a saturated model, ln L_S, is obtained by inserting $\frac{1}{n}$ for q_i in (23):

$$\ln L_\mathrm{S} = \frac{1}{n} \sum_{i=1}^{n} (-\ln q_i) = \frac{1}{n} \sum_{i=1}^{n} (-\ln \frac{1}{n}\ ) = \frac{1}{n} \cdot n \cdot \ln n = \ln \mathrm{n} \tag{36}$$

Furthermore, the null model is the model with all parameters $\theta_k = 0$ ($k = 1, ..., m$). The log loss of the null model, ln L_0, is obtained by inserting $\theta_k = 0$ into (29):

$$\ln L_0 = \ln \left(\sum_{i=1}^{N} e^{\Sigma_{k=1}^{m} \theta_k x_{ik}}\right) - \sum_{k=1}^{m} \theta_k \cdot \bar{x}_k^* = \ln N \tag{37}$$

Expressions (36) and (37) show that the log loss of the saturated model and the null model depend on the number of observed presences, n, and the total number of grid cells used for modelling, N, respectively, but that both are independent of the supplied environmental information $\mathbf{Z}$ and the variables $\mathbf{X}$ derived from the explanatory variables. For example 1* the log loss of the saturated MaxEnt model Q_S is ln L_S = ln 10 = 2.3036 while the null model Q_0 has ln L_0 = ln 40 = 3.6889.

Inserting $\theta_k = 0$ in (27) and then inserting for e^{θ_0} in (24) shows that the values of q_i under the null model are all equal to $\frac{1}{N}$, as desired:

$$e^{\theta_0} = \frac{1}{\Sigma_{i=1}^{N} e^{\Sigma_{k=1}^{m} \theta_k x_{ik}}} = \frac{1}{\Sigma_{i=1}^{N} e^{\Sigma_{k=1}^{m} 0}} = \frac{1}{\Sigma_{i=1}^{N} 1} = \frac{1}{N} \Leftrightarrow q_{0i} = e^{\theta_0 + 0} = e^{\theta_0} = \frac{1}{N} \tag{38}$$

The MaxEnt null model Q_0 thus predicts equal probability for a particular presence cell i_0, selected at random from all presence cells, to be cell i, for all $i = 1, ..., N$. Because $N > n$, ln $N >$ ln n. All MaxEnt models Q_t therefore have log loss values bounded below by ln n and above by ln N.

The maximum likelihood MaxEnt model for a specific set $\mathbf{X}_j$ of derived variables is the model Q_M with lowest log loss. The difference between ln N and ln n is a measure of the *total variation* in the response variable. This is an expression of the total variation possible to account for by a model. This measure is analogous with, but does not directly correspond to, the total deviance of GLM models. Furthermore, the strict maximum likelihood explanation of MaxEnt provides measures of variation accounted for, and not accounted for, by each specific MaxEnt model Q_t, defined in terms of the average log loss of observed presence grid cells. Three important statistics analogous with the explained deviance, the residual deviance and the fraction of explained deviance in GLM can be derived from (29), (36) and (37): the *variation accounted for* (VA) by model *t*, v_t, the *residual variation* of model t, w_t,and the *fraction of total variation accounted for* (FTVA) by model *t*, V_t:

$$v_t = \ln L_0 - \ln L_t v_t = \ln L_0 - \ln L_t \tag{39}$$

$$w_t = \ln L_0 - \ln L_\mathrm{S} - w_t = \ln L_0 - \ln L_\mathrm{S} - (\ln L_0 - \ln L_t) = \ln L_t - \ln L_\mathrm{S} \tag{40}$$

$$\mathrm{Vt} = \frac{\ln L_0 - \ln L_t}{\ln L_0 - \ln L_\mathrm{S}} \tag{41}$$

The MaxEnt model $Q_{1,\mathrm{L}}$ for example data set 1[*] (Fig. 2) with the L variable derived from explanatory variable 1 as the only derived variable (see Table 3) has parameters $\theta_0 = -2.3160$ and $\theta_1 = -4.7286$. Inserting for θ_0 and θ_1 in (12) we obtain:

$$q_i = e^{\theta_0 + \Sigma_{k=1}^{m} \theta_k x_{ik}} = e^{-2.3160 - 4.7286 \cdot x_{ik}} \tag{42}$$

Predictions from the one-variable MaxEnt model Q_1 (with $\boldsymbol{X}_1$ as the only derived variable) are $q(0.000) = 0.0987$, $q(0.143) = 0.0502$, $q(0.286) = 0.0256$, $q(0.429) = 0.0130$, $q(0.571) = 0.0066$, $q(0.714) = 0.0034$, $q(0.857) = 0.0017$ and $q(1.000) = 0.0009$. The q_i values sum to 1 over the 40 grid cells in $\boldsymbol{D}$. Log loss for the model is ln $L_1 = 2.9916$. From (39), (40) and (41) we obtain

$$v_1 = \ln L_0 - \ln L_t = 3.6889 - 2.9916 = 0.6973$$

$$w_1 = \ln L_t - \ln L_\mathrm{S} = 2.9913 - 2.3036 = 0.6877$$

$$\mathrm{V1} = \frac{\ln L_0 - \ln L_t}{\ln L_0 - \ln L_\mathrm{S}} = \frac{3.6889 - 2.9916}{3.6889 - 2.3036} = \frac{0.6973}{1.3853} = 0.5034$$

$V_1 = 0.5034$ means that 50.34 % of the total variation is accounted for by the model Q_1.

The MaxEnt model with the L variable derived from explanatory variable 2 as the only derived variable (see Table 3) has parameters $\theta_0 = 3.3228$ and$\theta_1 = -0.8142$.

Log loss for the model Q_2 (with $\boldsymbol{X}_2$ as the only derived variable) is $L_2 = 3.6485$, which corresponds to $v_1 = 0.0404$, $w_1 = 1.3449$ and $V_1 = 0.0300$.

Model selection (Step 8,i)

Model selection strategies

Finding the most parsimonious model, i.e., the model which best combines simplicity in terms of low number of parameters in the model and high possible predictive ability, and that circumvents overfitting, is a fundamental challenge of statistical modelling, distribution modelling included (Halvorsen 2012; also see 'Introduction' chapter). This task is accomplished by Step 8,i in the DM process, model selection, in which specific procedures are used to choose among alternative models. Note that a clear distinction is made here between model selection, which makes use of one and only one set of response variable data (the 'training data') together with explanatory data to find the parameterisation that, according to a set of criteria, is the best, and model evaluation, which implies assessment of model performance by use of data not *directly* used to parameterise the model. Thus, model selection is *internal* while model evaluation is *external* model performance assessment (Halvorsen 2012).

Hastie et al. (2009) recognise to two principally different model selection strategies:

1. *Subset selection methods* by which a discrete subset of derived variables is selected and the rest discarded, typically by omission of derived variables that do not contribute significantly to the predictive performance of the model. *Model comparisons*, i.e., assessment of the relative performance of two nested models, is central in subset selection. Two models are nested when the simpler model Q_{t-} is a submodel of the more complex model Q_t, i.e., that Q_{t-} only includes derived variables also included in Q_t. Subset selection is typically accomplished as a sequence of model comparisons; in each step applying a pre-selected method for internal model performance assessment and pre-selected model improvement criterion.
2. *Shrinkage methods*, by which the model coefficients θ_k are shrunk by imposing a penalty on their magnitude.

Both strategies make use of an *optimisation criterion* (OC), a performance index that penalises models for lack of fit to the data (bias) and model complexity. The general expression for a model optimisation criterion [Reikeking & Schröder (2006), expression (1)] is:

$$\text{OC} = (\text{model lack-of-fit}) + \lambda \cdot (\text{model complexity}) \qquad (43)$$

where λ is a user-defined regularisation parameter which determines the balance between model performance and model complexity. The better model is the one with the lowest OC value. Reineking & Schröder (2006) characterise λ as follows: 'One can think of this parameter as the exchange rate between the two 'currencies' of model lack-of-fit and model complexity.' The term regularisation is used here in the widest sense, for approaches that assist the trade-off between the fit of the model to data (reduction of bias) and model complexity (increase of prediction error) (cf. Reikeking & Schröder 2006). Subset selection methods penalise models for complexity by applying a function of the number of parameters while shrinkage methods trade complexity for bias by allowing more parameters, each of which with lower absolute value. Both approaches to model selection are captured by an alternative expression to (43) for the model optimisation criterion,

$$\text{OC} = (\text{model lack-of-fit}) + \lambda \cdot \sum_{k=1}^{m} |\,\theta_k|^{\zeta} \qquad (44)$$

where the parameter ζ determines the kind of model selection approach (Reineking & Schröder 2006). Subset selection corresponds to $\zeta = 0$; model complexity then only depends on the number of nonzero parameters in the model and the regularisation parameter λ. With $\zeta > 0$, models are penalised for high absolute values of parameters *in addition to* being penalised for high number of parameters. The penalty for large absolute values of the parameters increases with increasing value of ζ. Shrinkage methods represent a more continuous way to avoid overfitting because parameters θ_k associated with the derived variables $\boldsymbol{X}_k$ are shrunk (i.e., their absolute value reduced) rather than the derived variable as such omitted.

Subset selection

By subset selection, a subset of all variables derived from supplied explanatory variables by transformation, is selected. Subset selection can be carried out in many different ways, of which the theoretically optimal but in most cases undoable is the best subset method by which all possible combinations of subsets are compared. Several alternative methods exist for 'seeking a good path through' the innumerable combinations of derived variables (Hastie et al. 2009), but none come with a guarantee that the best subset is found. The choice among alternative subset selection methods therefore by and large remains a matter of personal preference. The two main pathways for subset selection are forward stepwise selection and backward stepwise elimination (e.g., Hastie et al. 2009). Forward selection starts with fitting the null model M_0, i.e., the model with the intercept θ_0 as the only model parameter. Thereafter, derived variables are added sequentially until a more complex model that performs relatively better than the best model found so far cannot be obtained. Backward elimination starts with the full model and sequentially leave out derived variables until no more derived variables can be omitted without reducing model performance. Backward elimination is impractical for DM unless the number of explanatory variables and, most notably, the number of variables derived from these variables, is low, but has been used in some studies (e.g., Lahoz-Monfort et al. 2007, Lamb et al. 2008, Parolo et al. 2008). A compromise between the two stepwise selection methods is the forward-backward procedure by which removal of derived variables is considered at the end of each step in the forward selection procedure. Forward-backward selection of variables has been used in several DM studies (e.g., van Niel & Austin 2007, Barbosa et al. 2009, Varela et al. 2009).

Opening for sets of variables to be derived from each explanatory variable adds considerable complexity to the subset selection procedure because a two-step procedure is required: (1) selection among variables derived from each explanatory variable, and (2) selection among explanatory variables, each represented by the set of derived variables resulting from (1).

A manual procedure for forward stepwise selection of derived variables and explanatory variables in MaxEnt is outlined in Table 4. Four main steps are recognised: (1) initial steps in which DVs are derived from EVs by transformation and methods and approaches for internal model performance assessment are specified: (2) analysis of single DVs; (3) selection of *parsimonious sets of DVs* for each EV; and (4) selection of MaxEnt model.

Shrinkage methods

Several shrinkage methods have been proposed, corresponding to different values for ζ in expression (44) (Reineking & Schröder 2006, Hastie et al. 2009). Choice of shrinkage method for MaxEnt is discussed by Dudík et al. (2007). Here I will restrict my attention to the shrinkage method implemented in the latest version 3.3.3k of the Maxent software (Phillips & Dudík 2008, Phillips et al. 2011), ℓ_1-regularisation, or lasso penalty (Tibshirani 1996), which corresponds to $\zeta = 1$.

In the context of MaxEnt modelling, ℓ_1-regularisation implies that model coefficients θ_k are found by minimising the penalised log loss, ln Λ_t, given by expression (45), instead of log loss, ln L_t, given by (29):

$$\ln \Lambda_t = \ln \mathrm{Lt} + \sum_{k=1}^{m} \lambda_k \cdot | \theta_k | = \ln \left(\sum_{i=1}^{N} e^{\Sigma_{k=1}^{m} \theta_k x_{ik}} \right) - \sum_{k=1}^{m} \theta_k \cdot \bar{x}_k^{\,*} + \sum_{k=1}^{m} \lambda_k \cdot | \theta_k | \qquad (45)$$

λ_k are regularisation parameters, i.e., constants set *a priori*, separately for each derived variable or type of derived variables. The model complexity penalty term $\Sigma_{\mathrm{k=1}}^{m} \lambda_k \cdot | \theta_k |$ in expression (45) is zero if all regularisation parameters $\lambda_k = 0$, which corresponds to regularisation being disabled, or if all $\theta_k = 0$, which implies that the model is reduced to the null model. Accordingly,

Table 4. Outline of a manual procedure for forward stepwise selection of derived variables (DVs) and explanatory variables (EVs) in MaxEnt (interaction DVs not specifically taken into account). ISDV = individually significant DV.

Step	Term	Description
1	**Initial steps**	
1a	Specification of method for internal model performance assessment	Select a method for comparing two nested MaxEnt models (e.g., the randomisation or *F*-ratio test, or ΔAUC), including a threshold (*model improvement criterion*; e.g., a significance level α or a ΔAUC value) to be used to decide if a more complex Maxent model is better than a simpler model
1b	Construction of continuous DVs	Transform each EV into continuous DVs, i.e., DVs of the L, M, and D types
1c	Construction of spline DVs	Transform each EV into variables of the spline type (HF, HR, T, and X). Because (almost) infinitely many DVs of each type can be constructed by shifting the position of the knot, a method for limiting the number of DVs is required.
2	**Analysis of single DVs**	
2a	Single-variable modelling step	For each DV, make a one-variable MaxEnt model without regularisation
2b	Single-variable test step	Compare each single-variable model with the MaxEnt null model by use of the method and the criterion specified in Step 1a
2c	Single-variable selection step	Select all ISDVs, i.e., DVs that satisfy the criterion specified in Step 1a, for use in Step 3, and leave out all other DVs and EVs for which no ISDVs could be obtained.
3	**Selection of parsimonious sets of DVs for each EV**	
3a	Finding the best DV in each set	For each set of ISDVs derived from the same EV, select the ISDV that performs best in the single-variable test of Step 2b. For sets with only one ISDV, this ISDV makes up the parsimonious set of DVs for this EV. For all other EVs, proceed to step 3b.
3b	Model improvement test step	For each additional ISDV in each set, make a two-variable MaxEnt model by adding this ISDV to the best ISDV in the set. Use the criterion specified in Step 1a to compare each two-variable model with the one-variable model for the best ISDV in the set.
3c	Set expansion step	Consider the following three cases: (i) If no ISDV is found in Step 3b that satisfies the criterion specified in Step 1a, the parsimonious set of DVs consists of the best ISDV only. (ii) If one ISDV satisfies the criterion, the parsimonious set consists of this ISDV and the best ISDV. (iii) If more than one DV satisfies the criterion, select the one which performs best in the two-variable test of Step 3b and repeat Step 3b by comparing three-variable models with the best two-variable model. Repeat the process until no more ISDVs can be added to the set.
4	**Selection of MaxEnt model**	
4a	EV test step	Compare MaxEnt models for each EV represented by a parsimonious set of ISDVs with the MaxEnt null model. Select the best EV according to the criterion specified in Step 1a.
4b	Model improvement test step	Similar to Step 3b, but applied to EVs represented by parsimonious sets of DVs instead of single DVs
4c	Model expansion step	Similar to Step 3c, but applied to EVs represented by parsimonious sets of DVs instead of single DVs. Optionally, interactions among already selected EVs can be considered for inclusion in the model together with the remaining EVs.
4d	Termination step	The best MaxEnt model is found when neither more EVs represented by parsimonious sets of DVs nor interactions among already selected EVs improve the model, as judged by the criterion specified in Step 1a, can be found

$\ln \Lambda_t > \ln L_t$ if not all $\lambda_k = 0$ or $\theta_k = 0$.

Phillips et al. (2006) explain ℓ_1-regularisation as a relaxation of (34), the condition that has to be satisfied by all MaxEnt models, that $\tilde{x}_k = \bar{x}_k^*$ for all derived variables k: the best ℓ_1-regularised model has the lowest penalised log loss within the bounds on $\tilde{x}_k$ given by

$$|\tilde{x}_k - \bar{x}_k^*| < \lambda_k \text{ (for all } k). \tag{46}$$

Equivalence of conditions (45) and (46) can be deduced from (46). Let us consider two cases, (i) that the mean of derived variable $\boldsymbol{X}_k$ for presence grid cells, $\bar{x}_k^*$, is lower than the mean of $\boldsymbol{X}_k$ for all cells, $\bar{x}_k$, and (ii) the converse. In case (i) we obtain:

$$\bar{x}_k^* < \bar{x}_k \Rightarrow \bar{x}_k^* < \tilde{x}_k \Rightarrow |\, x_k - \bar{x}_k^* \,| = \tilde{x}_k - \bar{x}_k^* \tag{47}$$

Inserting for (47) in (46) gives

$$|\, \tilde{x}_k - \bar{x}_k^* \,| < \lambda_k \Leftrightarrow \tilde{x}_k - \bar{x}_k^* < \lambda_k \Leftrightarrow \tilde{x}_k < \bar{x}_k^* + \lambda_k \tag{48}$$

Similarly, for case (ii) we obtain:

$$|\, \tilde{x}_k - \bar{x}_k^* \,| < \lambda_k \Leftrightarrow -(\tilde{x}_k - \bar{x}_k^*) < \lambda_k \Leftrightarrow \tilde{x}_k < \bar{x}_k^* - \lambda_k \tag{49}$$

Maximal shrinkage allowed under ℓ_1-regularisation corresponds to the situation by which

$$\tilde{x}_k = \begin{cases} \bar{x}_k^* + \lambda_k \text{ if } \bar{x}_k^* < \bar{x}_k \\ \bar{x}_k^* - \lambda_k \text{ if } \bar{x}_k^* > \bar{x}_k \end{cases}, \tag{50}$$

The property to be minimised under regularisation can be obtained from (29) by inserting (50) and using that θ_k is negative when $\bar{x}_k^* < \bar{x}_k$ and positive otherwise. Accordingly,

$$\ln \Lambda_t = \begin{cases} \ln \left(\sum_{i=1}^{N} e^{\sum_{k=1}^{m} \theta_k x_{ik}} \right) - \sum_{k=1}^{m} \theta_k \cdot (\bar{x}_k^* + \lambda_k) \text{ if } \bar{x}_k^* < \bar{x}_k \\ \ln \left(\sum_{i=1}^{N} e^{\sum_{k=1}^{m} \theta_k x_{ik}} \right) - \sum_{k=1}^{m} \theta_k \cdot (\bar{x}_k^* - \lambda_k) \text{ if } \bar{x}_k^* > \bar{x}_k \end{cases}$$

$$\ln \Lambda_t = \begin{cases} \ln \left(\sum_{i=1}^{N} e^{\sum_{k=1}^{m} \theta_k x_{ik}} \right) - \sum_{k=1}^{m} \theta_k \cdot \bar{x}_k^* - \sum_{k=1}^{m} \theta_k \cdot \lambda_k \text{ if } \bar{x}_k^* < \bar{x}_k \\ \ln \left(\sum_{i=1}^{N} e^{\sum_{k=1}^{m} \theta_k x_{ik}} \right) - \sum_{k=1}^{m} \theta_k \cdot \bar{x}_k^* + \sum_{k=1}^{m} \theta_k \cdot \lambda_k \text{ if } \bar{x}_k^* > \bar{x}_k \end{cases}$$

$$\ln \Lambda_t = \begin{cases} \ln \left(\sum_{i=1}^{N} e^{\sum_{k=1}^{m} \theta_k x_{ik}} \right) - \sum_{k=1}^{m} \theta_k \cdot \bar{x}_k^* + \sum_{k=1}^{m} |\theta_k| \cdot \lambda_k \text{ if } \bar{x}_k^* < \bar{x}_k \\ \ln \left(\sum_{i=1}^{N} e^{\sum_{k=1}^{m} \theta_k x_{ik}} \right) - \sum_{k=1}^{m} \theta_k \cdot \bar{x}_k^* + \sum_{k=1}^{m} |\theta_k| \cdot \lambda_k \text{ if } \bar{x}_k^* > \bar{x}_k \end{cases} \tag{51}$$

which equals (45).

In the Maxent software, the regularisation parameters λ_k are determined by

$$\lambda_k = \lambda_K \sqrt{\frac{\operatorname{var}(x_k^*)}{n}} \tag{52}$$

where λ_K is a 'tuning parameter' specific to each category of derived variables ('feature type') and var($\boldsymbol{X}_k^*$) is the variance of derived variable X_k over the n presence cells. Taking the square root of the variance and dividing by the square root of n makes the radical take the form of a standard error (Elith et al. 2011). The regularisation parameter λ_k therefore corresponds to a confidence interval, the width of which is determined by the variable-type specific constants λ_K.

A conservative attitude to model selection is implicit in regularisation by parameter shrinkage because the full potential offered by the data for predicting high q values in presence points is not utilised. The degree of regularisation imposed by standard settings for ℓ_1-regularisation in Maxent software is intermediate between a fully discrete and a fully continuous subset selection approach because derived variables are omitted (parameters = 0) if initial parameter estimates are very low. Because the absolute value of parameters is reduced, predictions q_i from the best MaxEnt model with ℓ_1-regularisation are always more conservative, i.e., closer to predictions by the null model, which are $q_0 = \frac{1}{N}$ for all i, than predictions from a corresponding model without regularisation. Accordingly, $\tilde{x}_k$ for the best model with regularisation is closer to the overall mean of $\boldsymbol{X}_k$ over all cells, $\bar{x}_k$, than the corresponding value for the model without regularisation, for which $\tilde{x}_k = \bar{x}_k^*$. Shrinkage methods thus reduce prediction error compared to subset selection methods by accepting higher bias (Reineking & Schröder 2006, Hastie et al. 2009).

Internal model performance assessment (Step 8,ii)

Internal model performance assessment in GLM and other maximum likelihood modelling methods is typically based on variation measures such as the sum of squares or deviance, or penalised versions thereof, such as penalised information statistics like AIC and BIC (cf. Hastie et al. 2009). When applied to modelling with binary response variables, these measures are used under the assumption that the data are of P/A type. Use with PO data therefore either implicitly implies that uninformed background observations are treated as pseudo-absences or that only observed presence observations are used in the calculations, as suggested for likelihoods by Warren et al. (2010). Although Phillips & Dudík (2008) and Warren et al. (2010) suggest that penalised information statistics for model comparison based upon measures of variation such as sums of squares or deviance, can be developed for P/A data. This has, however, hitherto not been done for MaxEnt. Phillips et al. (2006) describe MaxEnt as a '[less] mature a statistical method as GLM or GAM [with] fewer methods for estimating the amount of error in a prediction'. In this chapter I review standard methods and approaches for internal performance assessment of maximum likelihood models, and discuss their applicability to DM by MaxEnt.

The likelihood-ratio test

The strict maximum likelihood explanation of MaxEnt opens for model comparison by the *likelihood-ratio test* (e.g., Hastie et al. 2009). Let Q_{t-} and Q_t denote nested MaxEnt models; Q_{t-} being a submodel of Q_t. This means that all derived variables in Q_{t-} are also included in Q_t, and that Q_t contains one or more derived variables not included in Q_{t-}. The likelihood functions for observed presence observations for Q_{t-} and Q_t, $L_{t-,+}$ and $L_{t,+}$, are combined into the likelihood ratio, $LR_{t-,t}$, as follows:

$$LR_{t-,t} = \frac{L_{t,+}}{L_{t-,+}} \tag{53}$$

By use of the chi-squared approximation of the log-likelihood ratio (e.g., Hastie et al. 2009) and inserting (22) for the likelihood functions in (53), we obtain:

$$2 \cdot \ln LR_{t-,t} \sim \chi^2{}_{m_t - m_{t-}}$$

$$2 \cdot \ln \frac{L_{t,+}}{L_{t-,+}} \sim \chi^2{}_{m_t - m_{t-}}$$

$$2\ (\ln L_{t,+} - \ln L_{t-,+}) \sim \chi^2_{m_t - m_{t-}}$$

$$2\ (-n \ln L_t - (-n \ln L_{t-})) \sim \chi^2_{m_t - m_{t-}}$$

$$2n\ (\ln L_{t-} - \ln L_t) \sim \chi^2_{m_t - m_{t-}} \tag{54}$$

where $\chi^2_{m_t - m_{t-}}$ denotes the chi-square (χ^2) distribution with $m_t - m_{t-}$ degrees of freedom. m_t and m_{t-} denote the number of parameters in the respective models (the intercept θ_0 included). In terms of variation accounted for, given by (39), we obtain

$$2n\ ((\ln L_0 - v_{t-}) - (\ln L_0 - v_t)) \sim \chi^2_{m_t - m_{t-}}$$

$$2n\ (v_t - v_{t-}) \sim \chi^2_{m_t - m_{t-}}; \tag{55}$$

i.e., that $2n$ multiplied with the difference in variation accounted for by the two models approximates a chi-square distribution with degrees of freedom equal to the difference in number of parameters between the two models. In terms of residual variation of the models, given by (40), we obtain from (54)

$$2n\ ((w_{t-} - \ln L_S) - (w_t - \ln L_S)) \sim \chi^2_{m_t - m_{t-}}$$

$$2n\ ((w_{t-} - w_t) \sim \chi^2_{m_t - m_{t-}} \tag{56}$$

The sequential *F*-ratio test

The likelihood-ratio test can be applied to all pairs of nested models, including the saturated and null models. Using expression (56) for comparison of Q_t with the saturated model Q_S, and the fact the residual variation w_S of the latter is 0, we obtain

$$2\text{n (wt} - \text{wS)} \sim \chi^2_{m_t - m_{t-}}$$

$$2\text{nwt} \sim \chi^2_{\eta - m_t - 1} \tag{57}$$

where η denotes the appropriate degrees of freedom for the saturated model, i.e., the effective number of independent observations of the response variable. The term '–1' results from the definition of m_t is including the intercept θ_0. The ratio of two χ^2-squared distributions, normalised by the appropriate degrees of freedom, is *F* distributed (Myers et al. 2002). Accordingly, we obtain from (56) and (57) the *F* statistic for comparison of nested MaxEnt models Q_{t-} and Q_t.

$$F_{m_t - m_{t-},\, \eta - m_t - 1} = \frac{\dfrac{(w_t - w_{t-})}{(m_t - m_{t-})}}{\dfrac{w_t}{(\eta - m_t - 1)}} = \frac{(w_t - w_{t-}) \cdot (\eta - m_t - 1)}{w_t\,(m_t - m_{t-})} \tag{58}$$

This statistic follows the *F* distribution with $m_t - m_{t-}$ and $\eta - m_t - m_{t-}$ degrees of freedom. Accordingly, the F-*ratio test* for comparison of nested models, typically used to compare nested GLM models (e.g., Sokal & Rohlf 1995, Zuur et al. 2007), also applies to MaxEnt models. The *F*-ratio test is used, most often with a pre-selected significance level α, to evaluate the null hypothesis that the more complex model Q_t does not explain, or account for, significantly more variation than the simpler model.

At least three realistic alternatives exist for the appropriate value of the important parameter η in (58):

1. $\eta = n$; the number of observed presence observations. *A priori* arguments in favor of $\eta = n$ are: (i) that the most important parameter in the expression for variation accounted for (39), the model's log loss given by expression (29), is $\bar{x}_k^*$, the mean of derived variable X_k over the n presence sites; and (ii) that, in accordance with the opinion of several authors (e.g., Phillips et al. 2006, Elith et al. 2011) that MaxEnt is a presence-only rather than a presence–pseudo-absence modelling method, the values of the derived variable(s) for the observed presence observations (relative to a *static* background) are the *basic* determinants of the model.
2. $\eta = N$; the total number of observed presence + uninformed background observations. Arguments in favour of $\eta = N$ are: (i) that a model's log loss given by expression (29) is not *only* determined by the term $\Sigma_{k=1}^{m} \theta_k \cdot \bar{x}_k^*$ and hence, by the values of the derived variable(s) in observed presence cells, but also by the value of $\boldsymbol{X}_k$ in all N grid cells used in the analysis; and (ii) that the null model against which model performance is evaluated is the model which predicts equal probability of presence in all N grid cells. Argument (i) is motivated by the contribution of x_{ki} for all N grid cells i to the model parameter θ_0, given by expression (28).
3. $\eta = N - n$; the total number of uninformed background observations. The argument in favour of $\eta = N - n$ is that the scale on which log loss for a given model is expressed, is bounded above by ln N and below by ln n, so that the total variation possible to account for is ln N – ln n [expressions (36) and (37)]. $\eta = N - n$ thus accords with a view that the n observed presence grid cells serve as a 'given' reference with which the uninformed background cells are compared.

In the 'Worked examples' chapter I show, by comparing results of F-tests with different alternatives for η with Maxent runs on randomised data sets, that $\eta = N - n$ is likely to be the appropriate degrees of freedom for the residuals in the MaxEnt null model, in accordance with alternative (3). Inserting for η in expression (58) gives:

$$F_{m_t - m_{t-}, N-n-m_t-1} = \frac{\dfrac{(w_t - w_{t-})}{(m_t - m_{t-})}}{\dfrac{w_t}{(N - n - m_t - 1)}} = \frac{(w_t - w_{t-}) \cdot (N - n - m_t - 1)}{w_t (m_t - m_{t-})} \qquad (59)$$

The degrees of freedom are given by the number of parameters θ_k in the respective models. Each derived variable of the spline types is associated with one and not with two degrees of freedom because all values x_{ki} of a ranged spline derived variable (DV) are uniquely determined from z_{ki} once the position of the knot is fixed.

In terms of parameter values for the optimisation criterion given by (44), Reineking & Schröder (2006) show that the λ of the F-ratio test corresponds to the (1 – α)-quantile of the χ^2 distribution with 1 degree of freedom. This follows from (54): the difference in log likelihood between two models differing in one parameter only is asymptotically χ^2-distributed with one degree of freedom, under the null hypothesis that the value of that parameter is zero . Thus, $\alpha = 0.01$ corresponds to $\lambda = 6.635$, $\alpha = 0.05$ corresponds to $\lambda = 3.841$, $\alpha = 0.1$ corresponds to $\lambda = 2.706$ and $\alpha = 0.25$ corresponds to $\lambda = 1.383$.

Sequential F-ratio tests can be used to evaluate the contribution of: (1) one single DV (see Table 2); (2) one DV or a group of DVs, added to a model with other DVs derived from the same EV; (3) a set of DVs derived from the one EV; (4) a set of DVs derived from the same EV, added to sets of DVs derived from other EVs; and (5) one interaction DV between two or more

EVs already represented in the model, added to a model.

Penalised information statistics

The maximum likelihood explanation for MaxEnt opens for use of regularisation approaches based upon model optimisation criteria (OC) which use statistics of the penalised likelihood (PL) type, of the general form given by:

$$\text{PL} = -2 \cdot (\text{log-likelihood}) + \lambda \cdot (m + 1) \qquad (60)$$

where m is the number of parameters in the model. Penalised information statistics for model optimisation use the deviance [minus 2 × (the difference in log-likelihood between a model and the corresponding saturated model)] as a measure of model performance and the number of model parameters (plus one) to measure model complexity. With $\lambda = 2$, expression (60) becomes the AIC (Akaike's information criterion; Akaike 1973) as given by (cf. Crawley 2007: 353):

$$\text{AIC} = -2 \cdot \text{AIC} = -2 \cdot (\text{log-likelihood}) + 2(m + 1). \qquad (61)$$

With $\lambda = \ln \eta$ (where η is number of independent observations of the response variable; here tentatively set to $N - n$; see above), expression (60) becomes BIC (the Bayesian information criterion; Schwarz 1978), which penalises model complexity stronger than AIC for larger data sets ($\lambda > 2$ for $\eta \geq 8$).

The expression for AIC given by (61) is adapted to MaxEnt models Q_t by inserting (22) and (40) in (61):

$$\text{AIC}_t = -2 \cdot n \cdot (\ln L_t - \ln L_S) + 2(m + 1) = -2 \cdot n \cdot w_t + 2(m + 1) \qquad (62)$$

AIC, with $\lambda = 2$, corresponds to $\alpha = 0.157$ in a sequential F-ratio test.

For BIC, the following expression is obtained:

$$\text{BIC}_t = -2 \cdot n \cdot w_t + \ln (N - n)(m + 1) \qquad (63)$$

Randomisation tests

If realistic null models can be generated, e.g., by randomisation of the training data, a null-model approach to model comparison may be advantageous compared to the F-ratio test or penalised likelihoods because randomisation (permutation, or Monte Carlo) tests have fewer implicit assumptions. Randomisation tests imply that models Q_{t-} and Q_t are compared by randomising the targeted EV $\boldsymbol{Z}_j$ (i.e., the EV that is represented by DVs in Q_t but not in Q_{t-}) U times, for each randomisation deriving the appropriate new DVs from the randomised EVs, and finding the MaxEnt model that corresponds to these DVs. For each of the $U + 1$ MaxEnt models, Q_t and U models for randomisations of the relevant subset of $\boldsymbol{Z}_j$, $Q_{t,u}$, a test statistic such as the difference in variation accounted for by the model, v_t or $v_{t,u}$, and the variation accounted for by model Q_{t-}, v_{t-}, is recorded. A p-value for the randomisation test is obtained by counting the number of times, U_0, a randomised model performs better than the reference model Q_t:

$$\text{p} = \frac{(u_0 + 1)}{(u + 1)}\,. \qquad (64)$$

Because all DVs derived from the same explanatory variable make up a dependent variable set, the randomisation test cannot be applied to individual DVs derived from the EVs. Thus, (direct) randomisation tests are available for cases (1), (3) and (5) listed in the chapter 'The sequential *F*-test', but not for cases (2) and (5).

The area under the receiver operating curve (AUC)

Receiver operating characteristic (ROC) curve analysis is by far the most extensively used tool for assessment of the performance of distribution models, now encountered in almost every DM study (cf. Franklin 2009). ROC curve analysis was developed during World War II as a tool in signal processing, and is now used in many branches of science. Standard references for ROC curve analysis are Metz (1978), Hanley & McNeil (1982), Murphy & Winkler (1987), Fielding & Bell (1997) and Pearce & Ferrier (2000b); also see Phillips et al. (2006).

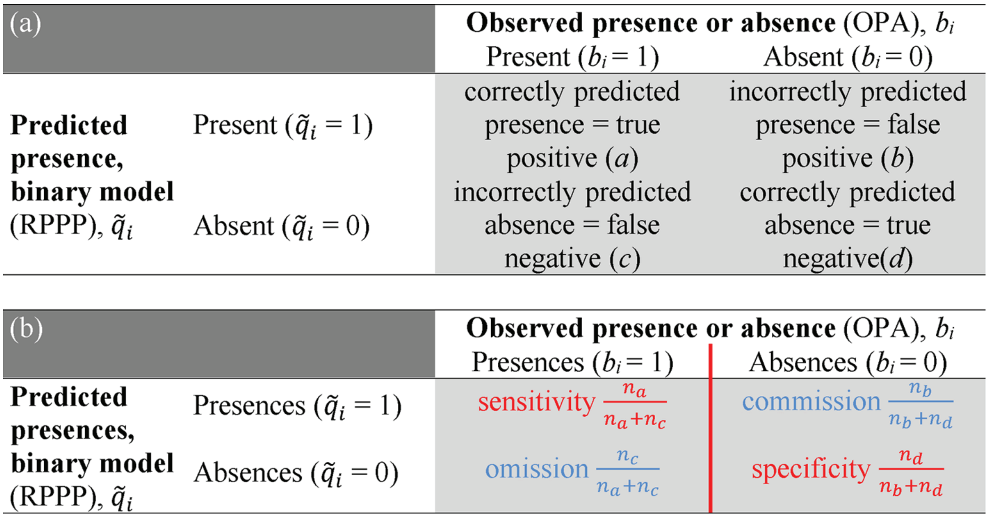
ROC curve analysis was originally devised to assess the performance of a model Q, the fitted values of which predicting the real probability of presence of a phenomenon for all instances of relevance to the study, by use of an independently collected evaluation data set D_e. In this original form, ROC curve analysis therefore applies to the model evaluation Step 11 in the 12-step DM process of Halvorsen (2012). However, ROC curve analysis can also be adapted to internal model performance assessment (Phillips et al. 2006). In this chapter, I first explain the basic principles of ROC curve analysis for model evaluation. Thereafter, I explain how ROC curve analysis can be adapted to internal model performance assessment of generative MaxEnt models.

Collection of an independent data set D_e of P/A observations of the modelled target, to be used for calibration and evaluation, is described as a separate Step 9 in the DM process (Halvorsen 2012). The set D_e can be described as follows: The 'instances of relevance' are sites d_i of unit size, i.e., grid cells. The set D_e contains N_e grid cells, selected to be representative for (but not necessarily a random sample of) all possible sites within the area of interest, which can be the study area or another area into which model predictions are to be transferred (PPM; Halvorsen 2012). For each site d_i, values for the observed presence or absence (OPA) vector B for the P/A data set are obtained; $b_i = 1$ means presence and $b_i = 0$ means absence. Note that the OPA vector B contains N_e elements, and that N_e is typically different from the number of grid cells, N, used for model parameterisation.

ROC curve analysis uses B together with model predictions q_i for the N_e observations in D_e. Any output format for MaxEnt model predictions that is monotonously related to the 'raw output', i.e., all five MaxEnt output formats described in the sections 'Output formats' and 'Model calibration and the probability-of-presence output format $\check{q}$', can be used for ROC curve analysis because this is a non-parametric statistical method. Only the ranks of the q_i are used in the computations.

The ROC curve is derived from confusion matrices, one for each unique value of q_i. The confusion matrices are obtained by a four-step process:

1. For each of the maximally $N_e - 1$ threshold values q_0, one in each interval between consecutively ordered values of q_i ($q_i < q_0 < q_{i+1}$), transform predictions from the continuous output scale (q_i or other output formats) to binary predictions $\tilde{Q}_{q_0} = (\tilde{q}_1, ..., \tilde{q}_i, ..., \tilde{q}_{N_e})$. Presence ($\tilde{q}_i = 1$) is predicted for $q_i \geq q_0$ and absence ($\tilde{q}_i = 0$) is predicted for $q_i < q_0$.
2. For each threshold value q_0 and each observation in the evaluation data set $\boldsymbol{D}_e$, make a decision matrix to record the appropriate combination of observed (b_i) and predicted ($\tilde{q}_i$) presence or absence [outcomes (a), (b), (c) and (d) in Fig. 3a].
3. For each threshold value q_0 construct a confusion matrix by counting the number of decision matrices with each of the four outcomes (a), (b), (c) and (d) in Fig. 3a; n_a, n_b, n_c and n_d.
4. From each confusion matrix (i.e., for each threshold value q_0), calculate the four



(a)		**Observed presence or absence** (OPA), b_i	
		Present ($b_i = 1$)	Absent ($b_i = 0$)
Predicted presence, binary model (RPPP), $\tilde{q}_i$	Present ($\tilde{q}_i = 1$)	correctly predicted presence = true positive (*a*)	incorrectly predicted presence = false positive (*b*)
	Absent ($\tilde{q}_i = 0$)	incorrectly predicted absence = false negative (*c*)	correctly predicted absence = true negative(*d*)

(b)		**Observed presence or absence** (OPA), b_i	
		Presences ($b_i = 1$)	Absences ($b_i = 0$)
Predicted presences, binary model (RPPP), $\tilde{q}_i$	Presences ($\tilde{q}_i = 1$)	sensitivity $\frac{n_a}{n_a+n_c}$	commission $\frac{n_b}{n_b+n_d}$
	Absences ($\tilde{q}_i = 0$)	omission $\frac{n_c}{n_a+n_c}$	specificity $\frac{n_d}{n_b+n_d}$

Fig. 3. Receiver operating characteristic (ROC) curve analysis by use of an independent presence/absence (P/A) evaluation data set D_e with N_e observations. (a) Decision matrix, showing the four possible combinations of observed (b_i) and predicted ($\tilde{q}_i$) presence or absence. Decision matrices are made for each combination of threshold value q_0 and observation i in D_e as an initial step in ROC curve analysis. A confusion matrix similar in shape to the decision matrix in (a) is obtained for each threshold value q_0 by counting the number of decision matrices with each of the four outcomes, n_a, n_b, n_c and n_d. (b) Performance statistics derived from the confusion matrix: fractions of cells with given observed presence status (observed presence, $n_a + n_c$, or observed absence, $n_b + n_d$; corresponding to the columns in the matrix as separated by the thick red line) that are correctly (red fonts) and incorrectly (blue fonts) predicted. Note that the two performance statistics in the same column sum to 1.

performance statistics (Fig. 3b):

$$\text{sensitivity} = \text{true positive rate} = \frac{n_a}{n_a + n_c} \tag{65}$$

$$\text{omission error} = \text{false positive rate} = \frac{n_b}{n_b + n_d} = 1 - \text{specificity} \tag{66}$$

$$\text{commission error} = \text{false negative rate} = \frac{n_c}{n_a + n_c} = 1 - \text{sensitivity} \tag{67}$$

$$\text{specificity} = \text{true negative rate} = \frac{n_d}{n_b + n_d} \tag{68}$$

The ROC plot shows corresponding values for omission error (horizontal axis) and sensitivity (vertical axes), one point for each unique threshold value. The ROC curve is the (broken) line that joins these points in order of increasing value for the threshold. The sensitivity and the omission error are independent of each other in the sense that the former indicates the model's ability to predict presence correctly while the latter expresses the model's tendency to predict presence incorrectly. The model with maximum possible predictive performance predicts presence and absence correctly for all sites in the evaluation data set. For this model, there exists a

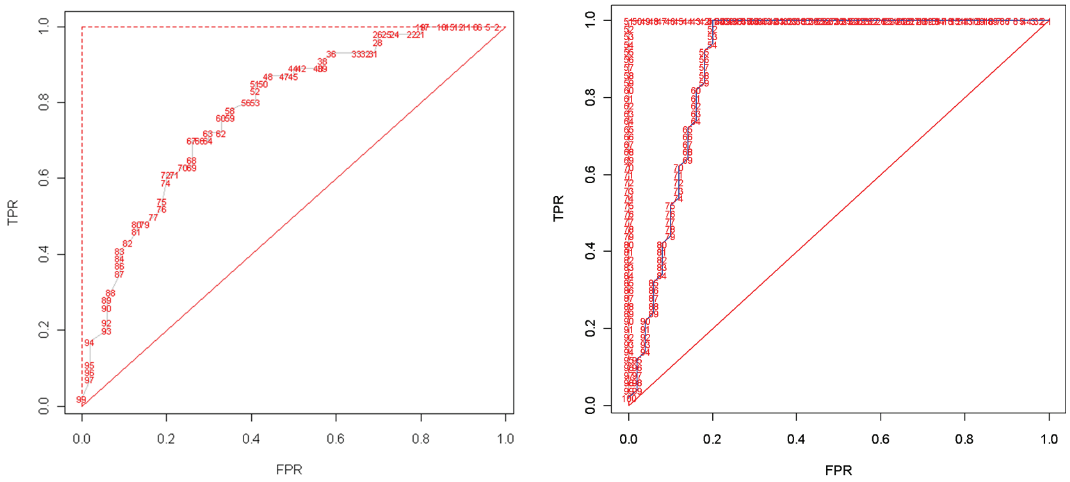


Fig. 4. Examples of ROC plots; FPR = false positive rate (omission error), TPR = true positive rate (sensitivity). (a) Typical ROC curve resulting from a model with 'fair' predictive ability (AUC = 0.781). The ROC curve starts in the upper right corner of the graph and joins points in order of increasing predicted value q_i (the red numbers are values of q_i × 100). The dotted red line from the upper right corner via the upper left corner to the lower left corner is the ROC curve for a model with maximal predictive power (AUC = 1.000). The continuous red line along the diagonal is the ROC curve for a random model (AUC = 0.500). (b) ROC curves for models with maximal predictive power, evaluated with P/A data (left curve; AUC = 1.000) and with PO data (right/lower curve; AUC = 0.900), respectively. The curve for PO data corresponds to a data set in which 20 % of the observations are observed presences.

threshold value (or a range of threshold values) q_0 for which $q_i \geq q_0 \Rightarrow b_i = 1$ (observed presence) and $q_i < q_0 \Rightarrow b_i = 0$ (observed absence). In the case exemplified by Fig. 3, the ROC curve runs from the upper right corner in the ROC plot via the upper left corner to the lower left corner (Fig. 4a). The ROC curve for a model that randomly assigns presence or absence to observations in the evaluation data set will tend to follow the diagonal (Fig. 4a) while models with predictive power between these extremes will have ROC curves somewhere in between. The area under the ROC curve, AUC, is therefore a measure of the predictive power of a model, judged over the entire range of possible threshold values. AUC can be calculated by the 'trapezoid method' (e.g., Pearce & Ferrier 2000b) by which the areas of all trapezoids under the ROC curve along the horizontal axis are summarised.

AUC values are often used for qualitative characterisation of distribution models. Perhaps the most frequently cited among such qualitative scales is a three-grade scale often referred back to Swets (1988). According to this scale, models are characterised as 'useful' if 0.7 < AUC ≤ 0.9 and as 'good' or 'excellent' if AUC > 0.9 (e.g., Kharouba et al. 2009, Nóbrega & de Marco 2011, Reside et al. 2011). However, in his original paper Swets (1988) advocates a more cautious and context-dependent interpretation of AUC values: 0.5 < AUC ≤ 0.7 as 'rather low accuracy'; 0.7 < AUC ≤ 0.9 as 'useful for some purposes'; and AUC > 0.9 as 'rather high accuracy'. A more explicit proposal of 0.7 and 0.9 as separating points on a three-grade scale was proposed by Pearce & Ferrier (2000b) who use the terms 'poor', 'reasonable', and 'very good'. Araújo et al.

(2005) expanded this scale to a five-grade scale on which 0.5 < AUC ≤ 0.6 is termed 'fail'; 0.6 < AUC ≤ 0.7 'poor'; 0.7 < AUC ≤ 0.8 'fair'; 0.8 < AUC ≤ 0.9 'good'; and AUC > 0.9 'excellent'. Many modified versions of these scales exist, and many authors have defined their own two-grade scale by defining a threshold AUC value to distinguish between 'random' and 'good' models; e.g., 0.6 (Trivedi et al. 2008, Parisien & Moritz 2011), 0.7 (Cordellier & Pfenninger 2009, Reside et al. 2011), 0.75 (Elith et al. 2006, Stachura-Skierczyńska et al. 2008), and 0.85 (Brown et al. 2008); see Merckx et al. (2011) for review.

Translations of AUC values into qualitative characterisations of model performance lack theoretical foundation (cf. Raes & ter Steege 2007, Nóbrega & de Marco 2011): what is a good and what is a poor model depends on data properties, the modelling purpose and, notably, the costs of erroneous predictions (Swets 1988). Furthermore, AUC is affected by data properties, e.g., sampling bias in the evaluation data (Raes & ter Steege 2007). The assumption that AUC is fully independent of the prevalence of the modelled target (e.g., Vaughan & Ormerod 2005, Raes & ter Steege 2007, Franklin 2009, Mouton et al. 2010) does not seem in general to hold true. In a recent study, Santika (2011) used simulated data to demonstrate that the relationship between AUC and prevalence is context dependent, influenced by the strength of the relationship between the modelled target and the dominant explanatory variable, the shape of the overall response curve with respect to this variable, and the degree to which the response is adequately modelled. Other studies in which relationships between AUC and prevalence have been addressed, include Luoto et al. (2005), Franklin et al. (2009), and Marmion et al. (2009).

When applied to P/A evaluation data, AUC expresses the probability that, if one presence cell and one absence cell are drawn at random from the pools of all presence and absence cells, respectively, the model will predict a higher RPPP value for the presence cell than for the absence cell (Hanley & McNeil 1982). The AUC is closely related to the Gini coefficient of size inequality (Gini 1912) and the non-parametric Wilcoxon-Mann-Whitney statistic (Phillips et al. 2006). More specifically, the area under an empirical ROC curve, calculated by the trapezoid method, is equal to the Wilcoxon-Mann-Whitney two-sample statistic applied to the two samples of presence and absence observations, respectively (DeLong et al. 1988). In case of ties (two or more evaluation grid cells with equal value of q, of which some are presences and some are absences), the AUC value will depend on the way ties are handled, i.e., the ordering of tied presence and absence observations, in the following way: AUC is lower if absence observations are placed before presence observations in the ordered list and *vice versa*. AUC values provided by Maxent software provides a balanced treatment of presences and absences in case of ties (cf. Phillips et al. 2006).

With PO data, e.g., the data used for parameterisation of a generative MaxEnt model, AUC can be calculated by replacing omission error with the probability that presence is predicted for an uninformed background cell picked at random. However, real presence cells are also likely to be included among uninformed background observations (Phillips et al. 2006). While with P/A data AUC is interpreted as the probability that the model predicts a higher RPPP value for a random presence grid cell than for a random absence cell, AUC with PO data is interpreted as the probability that the model predicts a higher RPPP value for an observed presence grid cell picked at random than for a randomly picked uninformed background cell (Phillips et al. 2006, 2009). The maximum achievable AUC value for PO data is not 1 but 1 – $\bar{C}/2$, where $\bar{C}$ is the frequency of observed presence of the modelled target in the set $\boldsymbol{D}$ of grid cells used for the study (Phillips et al. 2006). The explanation for this is that a fraction $\bar{C}$ of uninformed background cells, drawn randomly for comparison with observed presence cells, is expected to be real presence cells and that, by chance, one half of these cells is expected to have higher and one half is expected to have lower RPPP values than a random presence cell (cf. Fig. 4b). Similarly, the minimum achievable AUC value for a PO data set is not 0 but $\bar{C}/2$. PO-based AUC values ($\mathrm{AUC}_{\mathrm{PO}}$) calculated by the trapezoid method therefore have to be 'scale-corrected' to be comparable to AUC values for presence/absence data ('$\mathrm{AUC}_{\mathrm{PA}}$'):

$$\mathrm{AUC}_{\mathrm{corr}} = \frac{\mathrm{AUC}_{\mathrm{PO}} - \frac{\bar{C}}{2}}{1 - \bar{C}} \qquad (69)$$

Although $\mathrm{AUC}_{\mathrm{corr}}$ values are calibrated to a linear relationship with '$\mathrm{AUC}_{\mathrm{PA}}$', cases can be constructed for which $\mathrm{AUC}_{\mathrm{corr}}$ will exceed 1 (or be negative). The theoretically maximal value for $\mathrm{AUC}_{\mathrm{corr}}$ is obtained by inserting 1 for $\mathrm{AUC}_{\mathrm{PO}}$ in (69):

$$\mathrm{AUC}_{\mathrm{corr,max}} = \frac{1 - \frac{\bar{C}}{2}}{1 - \bar{C}} = \frac{1 - \bar{C}}{2(1 - \bar{C})} \qquad (70)$$

The difference in AUC between nested models t and t_-, the latter being a submodel of the former, can be used for internal model performance assessment:

$$\Delta\mathrm{AUC}_{t-1,t} = \mathrm{AUC}_{t} - \mathrm{AUC}_{t-1} \qquad (71)$$

The extent to which the areas under two ROC curves differ can, in principle, be tested by use of the general theory of Wilcoxon-Mann-Whitney *U*-statistics (e.g., DeLong et al. 1988), which allows the standard deviation of AUC values to be estimated [also see Hanley & McNeil (1982) and Pearce & Ferrier (2000b)]. However, confidence intervals constructed from these estimates are broad because the comparison between ROC curves relies on the implicit assumption that the two curves represent two random models rather than two nested models. Accordingly, this test is extremely conservative, with strong preference given to the simpler of the two nested models. Randomisation tests, by which (the) extra variable(s) of the more complex model is randomised *U* times and the number of times a randomised model has higher AUC than the model with the extra derived variable itself is counted, may be a good alternative, as suggested by Raes & ter Steege (2007) and applied by Reside et al. (2011). Such tests do, however, rely on the assumption that the same sampling bias is present in data used for evaluation and for parameterisation of the model. A *p* value for the test that the extra derived variable does not add significantly to the performance of the model, can obtained by expression (64).

Even though uncritical use of AUC has been rightfully criticised (e.g., Lobo et al. 2008), AUC has retained its position as a good overall indicator of model performance (e.g., Elith et al. 2006, Wisz et al. 2008).

Variable contribution to model (Step 8,iii)

Quantitative information about the relative contributions of single EVs, sets of DVs derived from one EV, or single DVs, to multi-variable MaxEnt models, are important properties of the parameterised model that results from Step 8,iii in the 12-step DM process. Variable contribution can be quantified in several ways. Four variable contribution measures are described here:

1. By *randomisation procedures*, i.e., by re-running the full model several times, each time randomising the variable the contribution of which is to be quantified. Performance reduction compared to the model in which variable is not randomised is recorded for each run. Performance reduction can be quantified by use of any statistic or approach outlined in the chapter 'Internal model performance assessment', e.g., the variation

accounted for, v_t, the fraction of total variation accounted for, V_t, or the AUC. The relative contribution of a variable is obtained as the ratio of the mean performance reduction resulting from randomisation of the variable in question and the sum of contributions of all variables. Contributions by interaction variables are distributed equally on the contributing variables. The term 'permutation importance' is used for this procedure as implemented in Maxent software with AUC reduction as performance statistic (Phillips 2011).

2. By *resampling of variables*, as exemplified by the procedure referred to as 'jackknifing' by Phillips (2011), by which the full model is re-run several times, each time leaving out one variable (an EV, a set of DVs derived from one EV, or a single DV). Performance reduction is recorded for each variable and the relative contribution calcutated as in (1). This procedure, with variation accounted for as performance statistic, is implemented in Maxent software as a graphical tool for assessment of variable contributions. The use of the term 'jackknifing' for this 'leave-one-*variable*-out' procedure is at odds with the normal use of the term, for 'leave-one-*observation*-out' resampling procedures (Sokal & Rohlf 1995).
3. By *null-model comparisons*, i.e., comparisons between single-variable MaxEnt models for all variables included in the full model with the null model. Examples of relevant performance statistics are the ΔAUC relative to the null model and the fraction of total variation accounted for (V_t) by single-variable models. A variable contribution statistic is obtained as the ratio of the contribution from the variable in question and the sum of contributions from all variables.
4. By *heuristic methods*, e.g., by recording for each step in the iteration process by which parameters of the final MaxEnt model are estimated [see Dudík et al. (2007) for explanation], the change, positive or negative, in variation accounted for resulting from changing the value of a model parameter θ_k. A relative measure of the contribution of each derived variable is obtained as the ratio of the sum of changes in variation accounted for by each DV, $\boldsymbol{X}_k$, and the total variation accounted for by the full model. This measure, which is implemented in Maxent software as 'percent contribution', is dependent on the path to the final model and is therefore regarded by Phillips (2011) as unreliable.

INTERPRETATION AND TRANSFORMATION OF MODEL PREDICTIONS (STEP 8,iv)

In this chapter I address interpretation of MaxEnt model predictions in geographical, environmental variables and ecological conceptual spaces. Three output formats, i.e., transformations of the raw output q, are described, of which two, the cumulative output and the logistic output, are implemented in the most recent version (3.3.3k) of the Maxent software (Phillips 2011).

Transferring predictions from discrete observation units in geographical and environmental variables spaces to continuous response functions

Generative MaxEnt distribution models are obtained by use of a data set $\boldsymbol{D} = \{d_1, ..., d_i, ..., d_N\}$, consisting of N observation units (grid cells) in abstract geographical space. Predictions $\boldsymbol{Q} = [q_1, ..., q_i, ..., q_N]^{\mathrm{T}}$ from this model are estimates of the *relative* probability of presence (RPPP) of the modelled target in each observation unit d_i. The vector $\boldsymbol{Q}$ can be represented as N points in discrete environmental variables space, the space in which points d_i are placed by their environmental characteristics vectors Z_i along axes defined by the explanatory variables $\boldsymbol{Z}_j$. Step 8 in the 12-step distribution modelling process, modelling of the overall ecological response, is completed when predictions for the N points are used to model a continuous response function, i.e., the overall ecological response in continuous environmental variables space. The translation from abstract geographical space via points in discrete environmental variables space to response curves in continuous environmental variables space is most easily explained by use of Bayesian statistical concepts (Phillips & Dudík 2008). In the chapter 'Outline of the MaxEnt statistical model', the quantity modelled by generative MaxEnt is denoted Pr $(i = i_0 | \, b_i = 1)$, while Pr $(b_i = 1| \, i = i_0)$, the real probability that the modelled target is present in a specific cell i_0, is characterised as the ideal output from distribution models. Recall at this point that Pr $(b_i = 1| \, i = i_0)$ cannot be confidently estimated if the prevalence of the modelled target is not known (Phillips et al. 2006, Ward et al. 2009).

Applying Bayes' rule to Pr $(b_i = 1| \, i = i_0)$, the relationship between this quantity and MaxEnt model 'raw output' Pr $(i = i_0| \, b_i = 1)$ is given as (Phillips & Dudík 2008):

$$\Pr\,(b_i = 1 \mid i = i_0) = \frac{\Pr\,(i = i_0 \mid b_i = 1) \cdot \Pr\,(b_i = 1)}{\Pr\,(i = i_0)}\,. \tag{72}$$

Here Pr $(i = i_0| \, b_i = 1)$ is the vector of MaxEnt estimates $\boldsymbol{Q} = [q_1, ..., q_i, ..., q_N]^{\mathrm{T}}$; Pr $(i = i_0)$ is the probability of picking grid cell i_0 *at random* from the set of all N grid cells in the study area, which is $\frac{1}{N}$ for all i; and Pr $(b_i = 1)$ is the prevalence of the modelled target, defined as the mean $\bar{b}$ of the P/A vector $\boldsymbol{B}$. Expression (72) can be simplified as follows:

$$\Pr\,(b_i = 1 \mid i = i_0) = \frac{q_i \cdot \bar{b}}{\frac{1}{N}} = N \cdot q_i \cdot b \tag{73}$$

Solving (73) for q_i we obtain

$$q_i = \frac{\Pr\,(b_i = 1 \mid i = i_0)}{N \cdot \bar{b}} \tag{74}$$

Expression (74) shows that the probability distribution $\boldsymbol{Q}$ estimated by MaxEnt is proportional to Pr $(b_i = 1| \, i = i_0)$, the real probability that the modelled target is present in a specific cell i_0 (Phillips & Dudik 2008), with $\frac{1}{N \cdot \bar{b}}$ as proportionality factor.

The MaxEnt prediction q_i for any grid cell d_i in $\boldsymbol{D}$ is obtained by inserting values for the explanatory variables Z_i or, if derived variables $\boldsymbol{X}_k$ have been obtained from $\boldsymbol{Z}_j$ by transformation, the values X_i for these derived variables obtained by the transformation function h using expression (2), into the parameterised Gibbs function applied in the MaxEnt model. This motivates for interpretation of the predictions $q(X)$ or, equivalently, $q(h(Z))$, as 'relative suitabilities' in environmental variables space, at least for all sites Z_l (or X_l) with environmental characteristics within the environmental range spanned by observations d_i. Almost all practical use of MaxEnt results for spatial prediction (Step 12 in the 12-step distribution modelling process) rests on such interpretation being valid.

The transfer from the MaxEnt distribution $\boldsymbol{Q}$ in abstract geographical space, expressed by (72), to environmental variables space occurs in two steps:

1. 'Translation' of predictions q_i for discrete points (observation units, grid cells) in abstract geographical space to points (grid cells) in discrete environmental variables space by use of the explanatory variable vectors Z_i or X_i for each of the N grid cells in $\boldsymbol{D}$ and the fact that $q_i = q(Z_i)$.
2. Generalisation from predictions for discrete points, $q(Z_i)$ to predictions for any site, $q(Z_l)$, in a subspace, a hypervolume, of the continuous environmental variables space [see Halvorsen (2012) for definition of conceptual spaces]. The subspace of interest is defined by the purpose of the DM study.

Calibration of MaxEnt models can be considered as a third step in the process by which raw predictions are 'transferred' into to estimates for Pr $(b_i = 1| \, i = i_0)$, the real probability that the modelled target is present in a specific cell i_0.

Step (1) starts with replacing '$i = i_0$', one specific grid cell in $\boldsymbol{D}$, with the set of environmental characteristics of this cell, X_{i0}, in expression (72):

$$\Pr\,(b_i = 1 \mid X_i = X_{i0}) = \frac{\Pr\,(X_i = X_{i0} \mid b_i = 1) \cdot \Pr\,(b_i = 1)}{\Pr\,(X_i = X_{i0})} = \frac{\Pr\,(X_i = X_{i0} \mid b_i = 1)}{\Pr\,(X_i = X_{i0})} \cdot \Pr\,(b_i = 1) \qquad (75)$$

The focus is thereby shifted from grid cells d_i *as such* to the environmental characteristics of these grid cells, expressed by the vector X_i: The probabilities on the right-hand side of (75) are interpreted as follows (Elith et al. 2011): Pr $(b_i = 1| \, X_i = X_{i0})$ is the probability that the modelled target is present in a grid cell with environmental characteristics given by the vector X_{i0} of values for the m DVs; Pr $(X_i = X_{i0}| \, b_i = 1)$ is the conditional probability that a presence grid cell has environmental characteristics given by the vector X_{i0}; and Pr $(X_i = X_{i0})$ is the unconditional probability that a grid cell has environmental characteristics X_{i0}. Pr $(X_i = X_{i0}| \, b_i = 1)$ can be estimated from PO data while Pr $(X_i = X_{i0})$ can be estimated from the set of all N grid cells in $\boldsymbol{D}$. As pointed out above, estimating the prevalence Pr $(b_i = 1)$ requires access to P/A data. Accordingly, access to P/A data allows Pr $(b_i = 1| \, X_i = X_{i0})$ to be modelled, as desired, while PO data supplied with environmental information for all grid cells allows modelling of $c \times$ Pr $(b_i = 1| \, X_i = X_{i0})$ where c is a scalar.

The next small step towards generalisation of predictions is the step from cell-specific probabilities given by (75) to probabilities for *sets* of grid cells with equal or similar environmental characteristics. Let X_f denote specifications for the vector of values for the m derived variables, and let n_f denote the number of grid cells in $\boldsymbol{D}$ with specifications complying with X_f. The elements x_f of X_f can be scalars or intervals. Furthermore, let X_{f0} denote one specific grid cell that complies with the specifications of X_f. Applying Bayes' rule to the set of grid cells with environmental characteristics complying with X_f, we obtain from (75):

$$\Pr\,(b_i = 1 \mid X_i = X_f) = \frac{\Pr\,(X_i = X_f \mid b_i = 1) \cdot \Pr\,(b_i = 1)}{\Pr\,(X_i = \mathrm{X}_f)} = \frac{\Pr\,(X_i = X_f \mid b_i = 1)}{\Pr\,(X_i = X_f)} \cdot \Pr\,(b_i = 1) \qquad (76)$$

Because, in general, Pr (A ∪ B) = Pr A + Pr B,

$$\Pr\,(b_i = 1 \mid X_i = X_f) = \textstyle\sum_{i:X_i \in \mathrm{X}_f} \Pr\,(X_i = X_{f0} \mid b_i = 1) = n_f \cdot \Pr\,(X_i = X_{f0} \mid b_i = 1) \qquad (77)$$

$$\Pr\,(X_i = X_f) = \textstyle\sum_{i:X_i \in \mathrm{X}_f} \Pr\,(X_i = X_{f0}) = n_f \cdot \Pr\,(X_i = X_{f0}) \qquad (78)$$

Solving (77) and (78) for n_f shows that the right sides of (75) and (76) are equal. Expression (73) shows that the relationships between the quotients $\frac{\Pr (X_i = X_0 \mid b_i = 1)}{\Pr (X_i = X_0)}$ and the MaxEnt predictions q are invariant of which environmental categories X_f are addressed:

$$\frac{\Pr (X_i = X_{i0} \mid b_i = 1)}{\Pr (X_i = X_{i0})} = \frac{\Pr (X_i = X_{f0} \mid b_i = 1)}{\Pr (X_i = X_{f0})} = N \cdot q(X) \Leftrightarrow q(X) = \frac{\Pr (X_i = X_f \mid b_i = 1)}{N \cdot \Pr (X_i = X_f)} \tag{79}$$

Rewriting expression (73), we obtain

$$\Pr (b_{i:X_i \in X_f} = 1 \mid X_{i:X_i \in X_f}) = \Pr (b_{i:X_i \in X_f} = 1 \mid X_i = X_f) = N \cdot q(X_f) \cdot \bar{b} \tag{80}$$

Expression (77) applies to all subsets $\boldsymbol{D}_f$ of grid cells in $\boldsymbol{D}$, regardless if the subset contains one single grid cell d_i, several grid cells with exactly the same environmental characteristics vector X_f, or many grid cells within a hypercube in environmental variables space. The quotient $\frac{\Pr (X_i = X_f \mid b_i = 1)}{\Pr (X_i = X_f)}$ given by expression (79), is termed the *presence-to-background frequency ratio,* is the ratio of the probability of encountering grid cells with environmental characteristics X_f in the subset of presence grid cells to the probability of encountering X_f in the set of all grid cells.

Pr $(X_i = X_f)$, Pr $(X_i = X_f \mid b_i = 1)$, and the presence-to-background frequency ratio can be illustrated by the L-type DV $\boldsymbol{X}_1$ in example data set 1* (Fig. 2, Table 3). It can be shown that a MaxEnt model for Sp with $\boldsymbol{X}_1$ as the only DV provides uniform predictions q for each level of $\boldsymbol{X}_1$ that are $q(0) = 0.0987$, $q(0.143) = 0.0502$, $q(0.286) = 0.0256$, $q(0.429) = 0.0130$, $q(0.571) = 0.0066$, $q(0.714) = 0.0034$, $q(0.857) = 0.0017$, and $q(1) = 0.0009$ (Table 5). Levels of $\boldsymbol{X}_1$ are uniform in number, thus Pr $(X_i = X_f) = \frac{1}{8}$ for each of the eight discrete values observed for the DV over the $N = 40$ grid cells. Of the $n = 10$ presence grid cells, four have $x_1 = 0$, three have $x_1 = 0.143$, two have $x_1 = 0.286$, one have $x_1 = 0.429$ while none have $x_1 = 0.571$, 0.714, 0.857, or 1. Accordingly, Pr $(X_1 = 0 \mid b_i = 1) = \frac{4}{10}$, Pr $(X_1 = 0.143 \mid b_i = 1) = \frac{3}{10}$, Pr $(X_1 = 0.286 \mid b_i = 1) = \frac{2}{10}$, Pr $(X_1 = 0.429 \mid b_i = 1) = \frac{1}{10}$, and Pr $(X_1 = (0.571 \vee 0.714 \vee 0.857 \vee 1.000) \mid b_i = 1) = 0$. According to expression

Table 5. Predictions from a MaxEnt model for the simulated target species Sp in example data set 1*, using the L-type derived variable (DV), $\boldsymbol{X}_1$, derived from explanatory variable (EV) $\boldsymbol{Z}_1$ in Table 3 as the only DV. Four different output formats for the predictions are given: q = raw output; $\dot{q}$ = probability-ratio output; $\ddot{q}$ = cumulative output; $\dddot{q}$ = logistic output, with two values for the logistic output parameter τ. Log loss of this MaxEnt model is 2.9915, which corresponds to a value of variation accounted for of $v = 0.6973$ and a fraction of variation accounted for of $V = 0.5030$.

Output format	τ value	X1 value							
		0.0000	0.1429	0.2857	0.4286	0.5714	0.7143	0.8571	1.0000
q	–	0.0987	0.0502	0.0256	0.0130	0.0066	0.0034	0.0017	0.0009
$\dot{q}$	–	3.9460	2.0070	1.0235	0.5197	0.2639	0.1359	0.0680	0.0360
$\ddot{q}$	–	0.4933	0.7441	0.8721	0.9370	0.9700	0.9870	0.9955	1.0000
$\dddot{q}$	0.25	0.3957	0.2499	0.1452	0.0794	0.0420	0.0221	0.0112	0.0059
$\dddot{q}$	0.50	0.6627	0.4998	0.3376	0.2056	0.1161	0.0634	0.0327	0.0176

(79), presence-to-background frequency ratios are $N \cdot q(X)$ both for individual grid cells d_i and for groups of grid cells with the same value for the derived variable so that, e.g., 40·0.0987 = 3.948 for $x_1 = 0.000$ and 40·0.0256 = 1.024 for $x_1 = 0.286$.

Step (2) generalises the interpretation of q_i as discrete probabilities $q(X_f)$ [or $q(Z_d)$] for single grid cells or grid-cell groups in discrete environmental variables space to an interpretation of $q(X_l)$ [or $q(Z_l)$] as a continuous response function, defined for all vectors X_l (or Z_l) in a hypervolume in the environmental variables space. Note the difference between vectors with subscripts *i*, *f*, and *l* here: the first refers to grid cells in abstract geographical space, the second to grid cells or grid-cell groups in discrete environmental variables space, and the third to points in continuous environmental variables space. The $q(X_i)$ and $q(X_f)$ differ fundamentally from the $q(X_l)$. While $q(X_i)$ can be interpreted as $(X_i = X_{i0} \mid b_i = 1)$ and $q(X_f)$ can be interpreted as $n_f \cdot \Pr(X_i = X_f \mid b_{i:X_i \in X_f} = 1)$, $q(X_l)$ at the outset lacks a specific meaning in environmental variables space because the constant *N* is a property of the data model, determined by the conceptualisation of the study area as a rasterised geographical space with *N* discrete grid cells, in Step 4 of the 12-step DM process. Secondly, note that for *arbitrary* points *i* in discrete environmental variables space, the probabilities which appear in expression (75), $\Pr(b_i = 1 \mid X_i = X_{i0})$, $\Pr(X_i = X_{i0} \mid b_i = 1)$, and $\Pr(X_i = X_{i0})$, and the corresponding probabilities subscripted *f* in expression (79) become probability density functions rather than discrete probabilities. Let $\Pr(b_i = 1 \mid X_i = X_l)$ denote the probability (density) that the modelled target is present at a site *l* (of unit size, equal to the size of grid cells d_i) with environmental conditions given by the vector X_l of values for the derived variables. The site *l* may correspond to a grid cell *i* in $\boldsymbol{D}$ or not. Let $\Pr(X_i = X_l \mid b_i = 1)$ denote the conditional probability density that a presence site has environmental conditions given by the vector X_l [$f_1(X_l)$ in the terminology of Elith et al. (2011); see Appendix 1], and let $\Pr(X_i = X_l)$ denote the unconditional probability density that an arbitrary site has environmental conditions given by the vector X_l [$f(X_l)$ in the terminology of Elith et al. (2011)]. Applying Bayes' rule to sets of grid cells with similar environmental characteristics, we obtain (81) as a parallel to (75) and (76):

$$\Pr(b_i = 1 \mid X_i = X_l) = \frac{\Pr(X_i = X_l \mid b_i = 1) \cdot \Pr(b_l = 1)}{\Pr(X_i = X_l)} = \frac{\Pr(X_i = X_l \mid b_i = 1)}{\Pr(X_i = X_l)} \cdot \Pr(b_l = 1) = \frac{\Pr(X_i = X_l \mid b_i = 1)}{\Pr(X_i = X_l)} \cdot \bar{b} \tag{81}$$

Expression (79) shows that the presence-to-background frequency ratio is modelled by $N \cdot q$ for all points that represent a subset $\boldsymbol{D}_f$ of grid cells X_f in environmental variables space. The continuous function of best fit to the *N* discrete values of $N \cdot q = \frac{f_1(X_l)}{f(X_l)}$ is obtained as the (smooth) hypersurface $N \cdot q(X_l)$ passing through the centroids of all grid-cell subsets. Thus, the quantity $N \cdot q(X_l)$ estimates the presence-to-background frequency ratio for all sites *l* [of unit (grid-cell) size] in the environmental variables space with values X_l for DVs $\boldsymbol{X}_k$:

$$\Pr(b_l = 1 \mid X_i = X_l) = \frac{\Pr(X_i = X_l \mid b_i = 1)}{\Pr(X_i = X_l)} \cdot \bar{b} \Leftrightarrow \frac{\Pr(X_i = X_l \mid b_i = 1)}{\Pr(X_i = X_l)} = N \cdot q(X_l) \tag{82}$$

Transferability of MaxEnt output from the set of *N* grid cells used to obtain the model to all grid cells of relevance for the purpose of the study is the reason why MaxEnt modelling can be performed by use of a restricted set of background observations ($\boldsymbol{D}$ instead of $\boldsymbol{D}_T$) without loss of generality of the results.

Elith et al. (2011) provide a different but equivalent explanation why MaxEnt predictions, given as presence-to-background frequency ratios, can be interpreted as a continuous response function in environmental variables space. From a machine-learning perspective, maximising the entropy of $\boldsymbol{Q}$, as defined in abstract geographical space, is equivalent to minimising the relative entropy of $f_1(X_l)/ f\,(X_l)$, i.e., the Kullback-Leibler divergence (Elith et al. 2011: Appendix S2). From this perspective, MaxEnt searches for the estimate of $f_1(X_l) = \Pr\,(X_i = X_l \mid b_l = 1)$, the probability density for the vector of derived variable values conditioned on presence of the modelled target, which best fits the presence data and which is (overall) as close to $f\,(X_l)$ as possible. In the terminology of Elith et al. (2011) referred to above, expression (81) becomes

$$\Pr\,(b_l = 1 \mid X_i = X_l) = \frac{f_1\,(X_l)}{f\,(X_l)} \cdot \Pr\,(b_l = 1) = \frac{f_1\,(X_l)}{f\,(X_l)} \cdot \bar{b} \tag{83}$$

The property of MaxEnt that its 'raw' output $\boldsymbol{Q}$ can be generalised from a discrete probability distribution in geographical space to relative predicted probabilities of presence in discrete environmental variables space and further to an overall ecological response in continuous environmental variables space by multiplication with N opens for the possibility that MaxEnt results may potentially be useful for ecological response modelling (ERM), in addition to spatial prediction modelling (SPM) and projective distribution modelling (PPM). In case of PPM, MaxEnt can be used both for projection into environmental and spatiotemporal scenarios different from the study area. In all relevant contexts, the fitted values $q(X_l)$ or $q(h(Z_l))$ from a MaxEnt model can be interpreted in environmental variables space as relative predicted probabilities of presence (RPPP) in a site of unit size with environmental characteristics vector Z_l.

Output formats

Raw output (q)

The term 'raw output', which is inherited from the Maxent software implementation of MaxEnt (Phillips et al. 2004, 2006, Phillips & Dudík 2008, Elith et al. 2011), is used for the model estimates $q(X_l)$ *as such*. The $q(X_l)$ are meaningful estimates of the *relative* predicted probability of presence (RPPP) of the modelled target in all sites l of unit size. However, because the sum of raw output values $q(X_l)$ over the N grid cells in $\boldsymbol{D}$ is 1, the average of $q(X_l)$ is $\bar{q} = \frac{1}{N}$ and $q(X_l)$ is therefore inversely proportional with the number N of background grid cells used in the modelling [expression (82)]. This context-dependence of the raw output restricts its *direct* relevance to situations where predictions are to be made for the specific set $\boldsymbol{D}$ of background grid cells d_i. In such cases q_i are estimates of the *probability* that *a specific* presence cell i_0, selected at random from all presence cells, is grid cell i; $\Pr(i = i_0 \mid b_{\mathrm{i}0} = 1)$. When a MaxEnt model is used to predict the RPPP in other observation units in $\boldsymbol{D}_T$, not included among background observations (the SPM purpose), or at sites outside the study area or under other scenarios (the PPM purpose), other output formats are therefore required. These other output formats may also be advantageous for prediction to $\boldsymbol{D}$ if they express properties of the modelled target that are more directly relevant to the purpose of the study than the raw output q_i. Values of q for the MaxEnt model for Sp with $\boldsymbol{X}_1$ as the only DV are given in Table 5.

Probability-ratio output ($\dot{q}$)

In the previous chapter, the presence-to-background frequency ratio was defined by expression (75) for the N grid cells in $\boldsymbol{D}$ and by expression (79) for subsets f of $\boldsymbol{D}$ with similar environmental characteristics. Both of these expressions show that q depends on N. A simple transformation from q to $\dot{q}$ by using expression (82), gives

$$\dot{q}_l = \Pr\,(b_l = 1 \mid X_i = X_l) = \frac{\Pr\,(X_i = X_l \mid b_i = 1)}{\Pr\,(X_i = X_l)} = N \cdot q(X_l) \tag{84}$$

The output format $\dot{q}$ removes the dependence of MaxEnt raw output q on the number N of background cells used in the modelling. $\dot{q}$ given by expression (84) can be interpreted as the 'presence-to-background frequency ratio', i.e., the ratio of the probability of encountering grid cells with environmental characteristics given by X_l in the subset of observed presence grid cells to the probability of encountering X_l in the set of all background grid cells. Elith et al. (2011) use the term 'relative suitability of one place vs. another' for $\dot{q}_l$ ($\frac{f_1(X_l)}{f(X_l)}$ in their terminology), and add that this quantity is 'giving insight about what features are important'. Hirzel et al. (2006) use the term 'predicted-to-expected ratio' for $\dot{q}_l$; also see Zielinski et al. (2010).

Values of $\dot{q}$ for the MaxEnt model for Sp with $\boldsymbol{X}_1$ as the only DV, given in Table 5, show that the presence-to-background frequency ratio predicted by MaxEnt varies from about 4 for $\boldsymbol{X}_1 = 0$, via 2 for $\boldsymbol{X}_1 = 0.1429$, to 0.0360 for $\boldsymbol{X}_1 = 1$. Values of $\boldsymbol{X}_1$ about 0.29 are predicted to have 'average suitability'.

$\dot{q}_l = 1$ means that the probability of presence of the modelled target in site l is estimated to be equal to the modelled target's probability of presence in a site chosen at random from the entire set of background cells. A benchmark for $\dot{q}_l$ is the probability ratio for presence sites in the saturated MaxEnt model, $q_{i,\mathrm{S}}$, which is obtained by inserting $q_i = \frac{1}{n}$ into (84):

$$\dot{q}_{i,\mathrm{S}} = N \cdot q_{i,\mathrm{S}} = N \cdot \frac{1}{n} = \frac{N}{n} \tag{85}$$

As will be demonstrated in the 'Worked examples' chapter, the ratio $\frac{N}{n}$ is, however, not an absolute maximum value for $\dot{q}_l$; neither for grid cells in $\boldsymbol{D}$ nor for sites with environmental characteristics outside the range of environmental variation spanned by background grid cells.

Elith et al. (2011: 46) state that the 'raw' output from Maxent software, i.e., q_l, are estimates of the ratio $\frac{\Pr\,(X_i = X_l \mid b_i = 1)}{\Pr\,(X_i = X_l)}$, i.e., of the presence-to-background frequency ratio. Expression (84) shows that is not the case and that, in order to be interpreted as a probability ratio, the raw output q_l has to be transformed to $\dot{q}_l$ by multiplication with N.

The interpretation of $\dot{q}$ as the ratio of probabilities of presence vs probabilities of encountering sites with a specific environmental characterisation has important implications for our understanding of sampling bias in the context of MaxEnt modelling. Phillips et al. (2009) point out that MaxEnt estimates q (and hence $\dot{q}$) are susceptible to sampling bias in the presence data. Let us assume that the grid cells in $\boldsymbol{D}$ provide accurate information about the unconditional probability $\Pr\,(X_i = X_l)$ of encountering sites with a specific environmental characterisation. This assumption will be met when all grid cells in the study area, or a random sample of these, are used as background cells [but see Lahoz-Monfort et al. (2007)]. The $\dot{q}(X_l)$ will be unbiased estimates of the relative suitability of a site X_l in environmental variables space if and only if the observed presence observations make up an unbiased sample from the population of all presence observations in the study area (or, more precisely, all grid cells in which the modelled target is present). Almost all PO data sets used for DM are, however, strongly biased (Kodric-Brown & Brown 1993, Vaughan & Ormerod 2003, Kadmon et al. 2004, Edwards et al. 2006, Hortal et al. 2008, Lobo 2008, Robertson et al. 2010, Wolmarans et al. 2010, McCarthy et al. 2011).

Sampling bias can be divided into geographical bias, i.e., overrepresentation of some, and underrepresentation of other parts of the study area in the data set compared to the areas occupied by these parts, and environmental bias, i.e., overrepresentation of some, and underrepresentation of other parts of the environmental variables space in the data set compared to their frequency in their study area (Wolmarans et al. 2010). The reliability of $\dot{q}$ as estimates of relative suitabilities rests on the extent to which the bias in the observed presence data $\boldsymbol{D}_+$ is 'outweighed' by similar sampling bias in the set $\boldsymbol{D}_-$ of uninformed background cells used in the modelling (Raes & ter Steege 2007, Elith et al. 2011). Subsets of uninformed background observations sampled to match the bias in presence observations, are referred to as 'target background' by Elith & Leathwich (2007); also see Phillips & Dudík (2008), Phillips et al. (2009) and Elith et al. (2011). Most experiments carried out so far indicate that target-group background observations improve the predictive power of MaxEnt models compared to models with random background observations (Phillips & Dudík 2008, Yates et al. 2010, Merckx et al. 2011; but see Loiselle et al. 2008).

Cumulative output ($\ddot{q}$)

The cumulative probability distribution corresponding to the raw output Q = $\{q_i\}$ is defined by

$$\ddot{q}_i = \sum\nolimits_{u:q_u \le q_i} q_u . \qquad (86)$$

The cumulative output format is available in Maxent software (Phillips et al. 2006) as $100 \cdot \ddot{q}$ (Phillips & Dudík 2008). Cumulative output expresses the probability that a specific presence cell i_0, selected at random from all presence cells, has a raw output value lower than q_i. The value q_i thus acts as a threshold value in the predictive context, separating the N grid cells into predicted presence cells ($q_u \ge q_i$) and predicted absence cells ($q_u < q_i$). Then, $\ddot{q}(q_i)$ gives the probability that absence is erroneously predicted for a randomly chosen observed presence cell. This quantity is the probability of erroneously predicting absence for an observed presence cell, i.e., the false negative rate, or omission rate [see the chapter ' The area under the receiver operating curve (AUC)']. Assignment of cumulative output values to sites l that are not grid cells in $\boldsymbol{D}$ rests on the assumption that the set of background grid cells is a representative sample from a hypervolume in environmental variables space that comprises site l.

Values of $\ddot{q}$ for the MaxEnt model for Sp with $\boldsymbol{X}_1$ as the only DV are given in Table 5.

Logistic output ($\dddot{q}$)

The raw output q, the probability-ratio output $\dot{q}$ and the cumulative output $\ddot{q}$ all express relative suitabilities in environmental variables space (cf. Elith et al. 2011), i.e., RPPP values or transformations thereof. Phillips & Dudík (2008) provide an output format (also see Dudík & Phillips 2009, Elith et al. 2011) claimed by them to be more intuitively interpretable: the *logistic output* $\dddot{q}$. The term refers to the 'logistic format' of q, i.e., that $\dddot{q}$ maps q onto a [0, 1] scale and thus resembles probabilities of presence, i.e., PPP values:

$$\dddot{q}(Z_l) = \frac{\tau q_l e^{\ln L}}{1 - \tau + \tau q_l e^{\ln L}} \qquad (87)$$

The logistic output parameter τ, which is to be chosen by the user, fixes the value $\dddot{q}(X^*)$for a site with average values for all derived variables x_{ik} over the n presence grid cells in $\boldsymbol{D}$, $X^* = (\bar{x}_1^{\,*}, ..., \bar{x}_k^{\,*}, ..., \bar{x}_m^{\,*})$. Elith et al. (2011) refer to X^* as 'an average presence site'. This interpretation of τ is confirmed by first inserting for X^* in (24) to obtain $q(X^*)$, followed by obtaining ln L from (29), multiplication of $q(X^*)$ with ln L, simplification of the product by use of the definition of θ_0 given by (28) and, finally, by inserting in (87):

$$q_i = e^{\theta_0 + \Sigma_{k=1}^{m} \theta_k x_{ik}} \Leftrightarrow q(X^*) = e^{\theta_0 + \Sigma_{k=1}^{m} \theta_k \bar{x}_k^*}$$

$$\ln L = \ln \left(\sum_{i=1}^{N} e^{\Sigma_{k=1}^{m} \theta_k x_{ik}}\right) - \sum_{k=1}^{m} \theta_k \cdot \bar{x}_k^*$$

$$q(X^*)\, e^{\ln L} = e^{\theta_0 + \Sigma_{k=1}^{m} \theta_k \bar{x}_k^* + \ln(\Sigma_{i=1}^{N} e^{\Sigma_{k=1}^{m} \theta_k x_{ik}}) - \Sigma_{k=1} \theta_k \bar{x}_k^*} = e^{\theta_0 + \ln(\Sigma_{i=1}^{N} e^{\Sigma_{k=1}^{m} \theta_k x_{ik}})}$$

$$q(X^*)\, e^{\ln L} = e^{\theta_0 - \theta_0} = 1$$

$$\dddot{q}(X^*) = \frac{\tau q(X^*) e^{\ln \mathrm{L}}}{1 - \tau + \tau q(X^*) e^{\ln \mathrm{L}}} = \frac{\tau}{1 - \tau + \tau} = \tau \qquad (88)$$

Logistic output is the default output format in recent versions of the Maxent software, among others recommended by Phillips & Dudík (2008), Elith et al. (2011) and Phillips (2011). This recommendation is motivated by the use of a probability (0–1) scale and by interpretability in terms of 'suitability' compared to an 'average presence site'. The resemblance of $\dddot{q}$ to real probabilities of presence, Pr (b_l = 1| X_l) is, however, only superficial because the 'logistic output at an average presence site', given by the logistic output parameter τ, does *not* correspond to the prevalence $\bar{b}$. With presence-only data, the choice of τ will *by necessity* be more or less arbitrary. In the absence of good reasons for choosing another value, Phillips & Dudík (2008) suggest τ = 0.5, which brings $\dddot{q}$ onto a 0–1 scale on which the value for an 'average presence site' is 0.5.

Values of $\dddot{q}_{0.25}$ and $\dddot{q}_{0.5}$, i.e., logistic output for values of τ of 0.25 and 0.5, respectively, for the MaxEnt model for Sp with $\boldsymbol{X}_1$ as the only DV, given in Table 5, show that the logistic output is less spread out than the raw output q. This is evident from the ratios $q(0)/q(1)$ = 109.67, $\dddot{q}_{0.25}(0)/\dddot{q}_{0.25}(1)$ = 66.66, and $\dddot{q}_{0.5}(0)/\dddot{q}_{0.5}(1)$ = 37.65. For both values of the parameter τ, the 'average presence site' is found to correspond to $\boldsymbol{X}_1$ = 0.143.

MODEL CALIBRATION AND THE PROBABILITY-OF-PRESENCE OUTPUT FORMAT $\breve{q}$ (STEP 10)

The logistic format $\dddot{q}$ brings MaxEnt output onto a probability-type of scale, resembling probabilities of presence (PPP), Pr (b_i = 1| X_i = X_l). This resemblance is, however, only superficial: unbiased estimates of Pr (b_i = 1| X_i = X_l) require explicit knowledge of the prevalence $\bar{b}$ of the modelled target in the study area or access to a P/A data set that can be used to estimate $\bar{b}$. Nevertheless, the recommendation of the logistic format by Phillips and co-workers demonstrates the importance of well-calibrated models for practical use of DM results, as also emphasised, among others, by Pearce & Boyce (2006), Reikeking & Schröder (2006) and Gastón & García-Viñas (2010). The extent to which a model is well calibrated can be inspected on a calibration plot (e.g., Pearce & Ferrier 2000b: Fig. 3, Edwards et al. 2005: Fig. 3, Edvardsen et al. 2011: Fig. 3). The calibration plot is a graph with the mean or median RPPP value, $\bar{q}_u$, for each class u into which the range of predicted probabilities is divided, on the horizontal axis, and the frequency of presence (FP), $\bar{b}_u$, i.e., the frequency of presence sites in each class, calculated from the P/A data set (Halvorsen 2012), on the vertical axis. A well-calibrated model is characterised by corresponding values for $\bar{q}_u$ and $\bar{b}_u$ that are close to the line $\bar{b}_u = \bar{q}_u$ (Pearce & Ferrier 2000b). A confidence interval for each $\bar{b}_u$ can be obtained by considering the set of N_u evaluation points in interval u as N_u binomial trials, each with probability $\bar{b}_u$ (Edvardsen et al. 2011).

Insertion of (82) into (81) shows that both of the raw output format q and the probability-ratio output format $\dot{q}$ are expected to be linearly related to the frequency of presence (FP):

$$\Pr\,(b_l = 1 \mid X_i = X_l\,) = N \cdot q(X_l) \cdot \bar{b} = q_l \cdot \bar{b} \qquad (89)$$

According to expression (89), the expectation of MaxEnt output is good calibration to probabilities of presence (PPP). In practice, however, this is not the case. Because $N \cdot q(X_l)$ is not bounded above by $\frac{N}{n}$, $N \cdot q(Xl\,) \cdot \bar{b}$ is not bounded above by 1. Use of (89) to calibrate MaxEnt output is therefore inappropriate for the same reason that ordinary linear regression models (LM) are inappropriate for modelling response variables of the probability type (e.g., Crawley 2007). In regression, this problem is resolved by customary use of GLM with logit link function and binomial errors (logistic regression) instead of LM. This motivates for similar measures to be taken if transformation of MaxEnt output to a probability ([0,1]) scale is required.

Pearce & Ferrier (2000b) provide a detailed review of DM calibration, among others showing how estimates for PPP, here termed probability-of-presence output and denoted $\check{q}_l$, can be obtained from RPPP on the q_l or $\dot{q}_l$ formats by use of independent evaluation data. Consider a P/A evaluation data set B_e which consists of N_e observation units of similar size as grid cells in the PO data set $\boldsymbol{D}$. Assume that RPPP estimates from a MaxEnt model and an observed presence or absence (OPA) vector, B_e, are available for all N_e grid cells. If the evaluation data set is a random sample of P/A observations of the modelled target, prevalence can be estimated directly as the frequency of presence in the evaluation data set (Halvorsen 2012). However, any stratified random sample of grid cells for which P/A data are obtained, e.g., by use of MaxEnt model output q_l and $\dot{q}_l$ for stratification (Edvardsen et al. 2011), can be used for model calibration. This makes collection of appropriate calibration and evaluation data possible also for species and other targets of DM that are too rare for random sampling to be practically and economically feasible (e.g., Phillips & Elith 2010). Transformation of q_l into $\check{q}_l$ is performed by fitting a calibration model to the RPPP–OPA relationship. Pearce & Ferrier (2000b) show that b_i can be appropriately modelled as the response to q_l (or $\dot{q}_l$) by a logit-logit relationship, which ensures that both the RPPP values, the response b_i, and the fitted values of the model, $\check{q}i$, are expressed on probability ([0,1]) scales:

$$\ln \frac{b_i}{1 - b_i} = \beta \cdot \ln \frac{q_i}{1 - q_i} + \beta_0$$

Values for b_l fitted by this model are the targeted probability-of-presence output $\check{q}_l$ for arbitrary sites l of unit grain size. $\check{q}_l$ is obtained by back-transformation, i.e., by solving (90) for $\check{q}_l$:

$$\ln \frac{\check{q}_l}{1 - \check{q}_l} = \beta \cdot \ln \frac{\dot{q}_l}{1 - \dot{q}_i} + \beta_0$$

$$\frac{\check{q}_l}{1 - \check{q}_l} = e^{\beta \cdot \ln \frac{\dot{q}_l}{1 - \dot{q}_i} + \beta_0}$$

$$\check{q}_l = \frac{e^{\beta \cdot \ln \frac{\dot{q}_l}{1 - \dot{q}_i} + \beta_0}}{1 + e^{\beta \cdot \ln \frac{\dot{q}_l}{1 - \dot{q}_i} + \beta_0}} \qquad (91)$$

The parameters β (the slope) and β_0 (the intercept) of the model given by (90) can be interpreted as the spread and the bias, respectively, of the RPPP-OPA relationship (Pearce & Ferrier 2000b). Methods also exist for testing the hypotheses that β_0 does not deviate from the expected value of 0 (Miller et al. 1991), and that β does not deviate from some scalar value.

SOME CONSIDERATIONS FOR EVALUATION OF MAXENT MODELS (Step 11)

Since procedures for model evaluation are generally applicable to distribution models regardless of choice of modelling method, only a few considerations that specifically apply to MaxEnt models will be included here. The reader is referred to Halvorsen (2012) for a full overview of the many good reasons that exist for evaluating distribution models by independently collected presence/absence (P/A) data

The obvious choice of performance statistic for model evaluation by independent P/A data is the prediction error, PE (Hastie et al. 2009), obtained as the sum of (squared) PEs for each observation in the evaluation data set. A logical choice of MaxEnt output format for calculation of PE is, in my opinion, the probability-of-presence output $\breve{q}$ given by expression (91). Alternatively, the logistic output format $\ddddot{q}$ may be used, for each model with the logistic output parameter τ chosen to minimise prediction error (E. Heegaard, personal communication).

The most frequently used performance statistic for assessment of the performance of distribution models is the AUC, which is explained in the chapter ‘ The area under the receiver operating curve (AUC)’. AUC is also applicable to the three strategies for model evaluation recognised by Halvorsen (2012), in addition to evaluation by independent P/A data, which use PO data: evaluation by data-splitting, evaluation by data resubstitution, and evaluation by repeated resubstitution of data. PE is not available as performance statistic when true absence observations are missing.

MAXENT MODELLING WITH PRESENCE/ABSENCE DATA

Use of P/A data for modelling the overall ecological response (Step 8)

P/A data, such as the independent P/A data collected for model calibration and evaluation (Step 9 in the 12-step DM process) can also be used in Step 8 to assist modelling of the overall ecological response of the modelled target. This is accomplished as follows: Firstly, a method for *external* model performance assessment is selected as a replacement for the internal model performance method of Step 8,ii. Prediction error (PE) and AUC on independent P/A data, which are unavailable for generative MaxEnt modelling with PO data, are eligible. Secondly, model selection (Step 8,ii) is performed by one of the procedures described in the chapter ‘Model selection’, using the replacement performance assessment method.

When P/A data are used in Step 8 of the modelling process, statistical tests such as the likelihood-ratio test, the *F*-ratio test, or randomisation tests, do not, at the outset, appear to be needed because the model’s ability to balance model fit and model complexity is expressed directly by the performance statistic. However, the probability that addition of a random variable gives rise to a model with a slightly higher value for the performance statistic increases with increasing number of EVs, or DVs derived from these EVs, to be tested. This is referred to as the multiple testing problem by Legendre & Legendre (1998). Randomisation tests may therefore be useful also when independent P/A data are used in Step 8, e.g., to determine a minimum threshold value for change of the performance statistic which corresponds to a pre-set significance level, e.g., $\alpha = 0.05$, in the randomisation test. As a shortcut, a pre-defined threshold ΔAUC value can be used as model improvement criterion, i.e., to assess if a more complex model is better than a simpler model. Choice of such a threshold ΔAUC value should be guided by data-set properties and previous experience with other data sets or, preferably, by randomisation tests on the data in question. Empirical evidence from worked examples will certainly be useful. Unpublished results of I. Auestad et al., (in prep.) suggest that ΔAUC values in the range 0.005–0.010 may be reasonable. In the absence of a ΔAUC value that can be reasonably argued for, alternative models should be obtained by use of different ΔAUC threshold values, and the resulting models compared.

A note on discriminative MaxEnt models

In a DM context, MaxEnt is almost exclusively used with PO data [but see Wollan et al. (2011)]. The MaxEnt method does, however, also apply to P/A data. Predictions from discriminative MaxEnt models, i.e., MaxEnt models parameterised by use of a P/A response variable, have a different interpretation than predictions from generative MaxEnt models. Furthermore, since P/A data are usually obtained by systematic or random sampling, response data for discriminative MaxEnt modelling are likely to have low bias compared with response data for generative MaxEnt modelling.

The response variable ($\boldsymbol{Y}$) used in MaxEnt modelling is the same in generative and discriminative MaxEnt: the *probability* that *one specific* presence cell i_0, selected at random from all presence cells, is grid cell i; $\Pr(i = i_0 \mid b_{i0} = 1)$. However, when $\boldsymbol{Y}$ is the vector $\boldsymbol{B}$ of observed presences or absences of the modelled target rather than the observed presence (OP) vector $\boldsymbol{C}$, the frequency of presence grid cells is an unbiased estimate for the prevalence of the modelled target:

$$b = \frac{n}{N}\ . \tag{92}$$

The raw (q), probability ratio ($\dot{q}$) and cumulative ($\ddot{q}$) output formats have the same interpretation in discriminative as in generative MaxEnt. However, insertion of (92) into expression (89) shows that with P/A data, an unbiased estimate for the probability-of-presence output is obtained directly as

$$\Pr\,(b_l = 1 \mid X_i = X_l\,) = N \cdot q_l \cdot \bar{b} = N \cdot q_l \cdot \frac{n}{N} = q_l \cdot n. \tag{93}$$

However, like in generative MaxEnt models, $\Pr(b_l = 1|X_l)$ given by (93) is not bounded above by 1 because q_l is not bounded above by $\frac{1}{n}$ (see the chapter ‘Probability-ratio output’). Calibration of model output to a [0, 1] scale is therefore required, This can be accomplished by fitting of the logit-logit function given by expression (90); see the chapter ‘Model calibration and the probability-of-presence output format $\check{q}$’.

All model selection methods and all methods and approaches for internal model performance assessment that apply to generative MaxEnt models also apply to discriminative MaxEnt models.

WORKED EXAMPLES

MATERIAL: SIMULATED DATA SETS

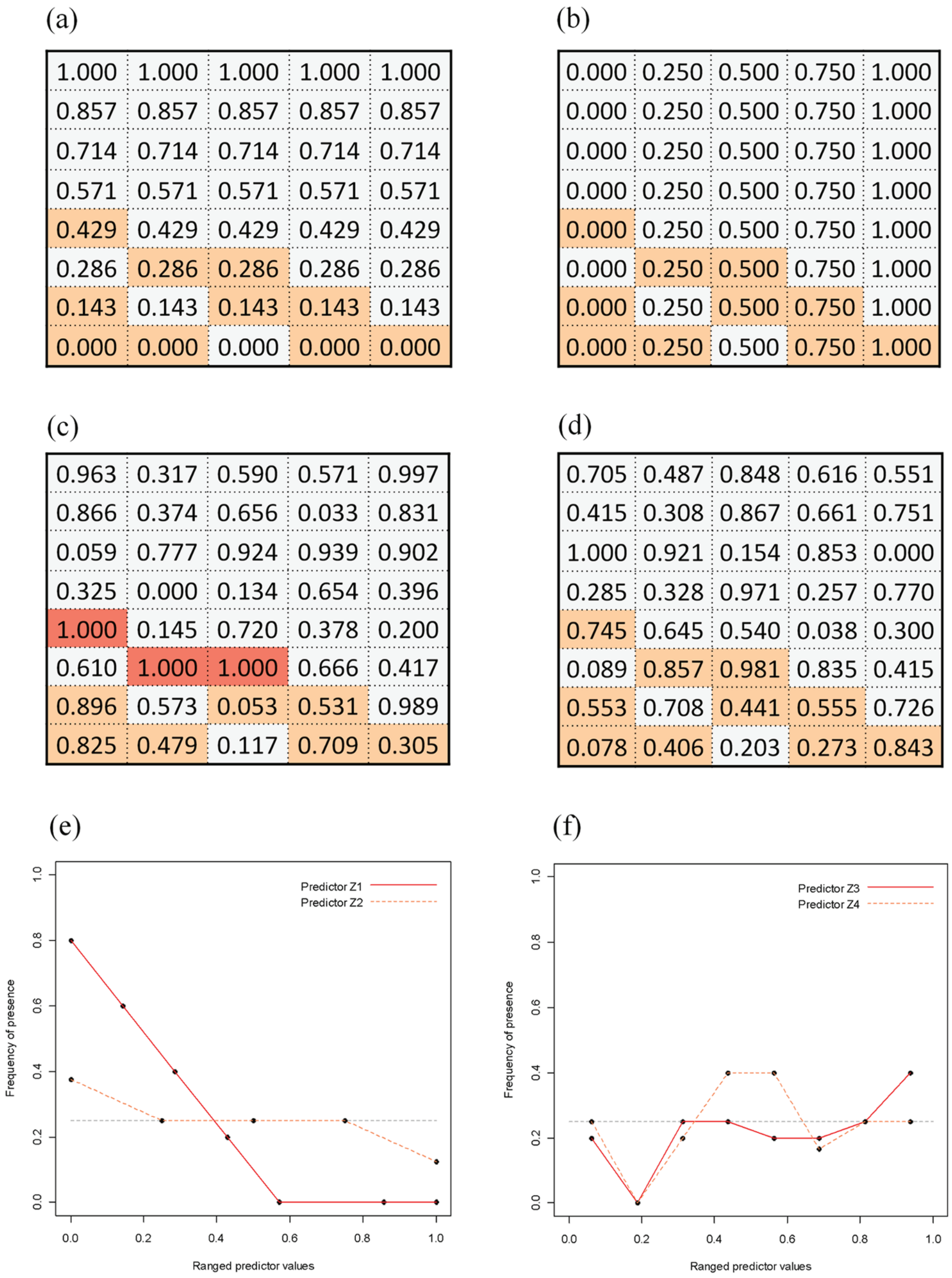
Two simulated data sets are used for the worked examples.

Example data set 1

Example data set 1 is similar to example data set 1* except for the addition of two explanatory variables. The study area is rasterised into 40 grid cells, arranged in 8 rows × 5 columns (Fig. 2a). The set of observation units is denoted $\boldsymbol{D}_1 = \{d_{1,1}, ..., d_{1,i}, ..., d_{1,40}\}$. A simulated target species 'Sp1' (= Sp in example 1*) is observed in $n = 10$ (25 %) of the $N = 40$ grid cells in $\boldsymbol{D}$ (Figs 2a, 5a). No information is available about eventual presence or absence of Sp1 in the remaining $N - n = 30$ uninformed background grid cells.

The environmental data set $\mathbf{Z}_1$ consists of four explanatory variables, $\boldsymbol{Z}_{1,j}$ ($j = 1, ..., s$; $s = 4$), each recorded for every grid cells in $\boldsymbol{D}_1$; $\boldsymbol{Z}_{1,j} = [z_{1,1j}, ..., z_{1,ij}, ..., z_{1,40j}]^{\mathrm{T}}$. $\boldsymbol{Z}_{1,1}$ indexes northing ('Y coordinate') in the rasterised geographical space representation of the study area (Fig. 2b) while $\boldsymbol{Z}_{1,2}$ indexes easting ('X coordinate') in this space (Fig. 2c). $\boldsymbol{Z}_{1,3}$ and $\boldsymbol{Z}_{1,4}$ are obtained as vectors of random numbers, drawn from a uniform distribution [0, 1]; $\boldsymbol{Z}_{1,3}$ modified so that three randomly chosen presence cells were given the maximum value of 1. The convention for sorting and indexing of grid cells described in the chapter 'Theory: Data sets' is adopted: the observed presence subset $\boldsymbol{D}_{1+}$, $i = 1$–10, contains observed presence grid cells, while $i = 11$–40 are uninformed background grid cells. Within each subset, grid cells are numbered consecutively by columns within rows from the 'SW' corner. Thus, the five grid cells in the lowermost row are indexed 1, 2, 11, 3 and 4, respectively, and the five grid cells in row 4 from below are indexed 10 and 17–20, respectively. The observed presence (OP) vector $\boldsymbol{C}_1 = [c_{1,1}, ..., c_{1,i}, ..., c_{1,40}]^{\mathrm{T}}$ contains information about observed presence ($c_{1,i} = 1$) or unknown presence or absence status ($c_{1,i} = 0$). Explanatory variables $\boldsymbol{Z}_{1,1}$ to $\boldsymbol{Z}_{1,4}$ were ranged, i.e., linearly rescaled to a [0,1] scale, to obtain L-type variables $\boldsymbol{X}_{1,1\mathrm{L}}$ to $\boldsymbol{X}_{1,4\mathrm{L}}$ (see Figs 5a–d). $\boldsymbol{X}_{1,1\mathrm{L}}$ and $\boldsymbol{X}_{1,2\mathrm{L}}$ are discrete variables with values that make up closed arithmetic sequences starting at 0 and ending at 1, with steps of $\frac{1}{4}$ and $\frac{1}{7}$, respectively (Figs 5a–b). Pr ($x_{1,1\mathrm{L},i} = \frac{c}{7}$) = 0.125 for all integers $c \in [0,7]$ and Pr ($z_{1,2\mathrm{L},i} = \frac{c}{4}$) = 0.2 for all integers $c \in [0,4]$.

The frequency of observed presence (Halvorsen 2012) of Sp1 with respect to derived variable $\boldsymbol{X}_{1,1\mathrm{L}}$ decreases markedly with increasing value of $\boldsymbol{X}_{1,1\mathrm{L}}$ from 0.8 at $x_{1,1\mathrm{L}} = 0$ to 0 at $x_{1,1\mathrm{L}} \geq \frac{4}{7}$ (Fig. 5e). The frequency of observed presence with respect to $\boldsymbol{X}_{1,2\mathrm{L}}$ varies but little, from 0.375 at $x_{1,2\mathrm{L}} = 0$ to 0.125 at $x_{1,2\mathrm{L}} = 1$. No clear patterns of variation in frequency of observed presence was found with respect to $\boldsymbol{X}_{1,3\mathrm{L}}$ and $\boldsymbol{X}_{1,4\mathrm{L}}$ (Fig. 5f). Frequency of observed presence is an important data-set property because estimates from MaxEnt, given as the probability-ratio output format $\dot{q}$, express as the ratio of the probability that the modelled target is present in a site with vector of values for explanatory variables Z_i (here represented by derived variables X_i) and the probability that a site is characterised by Z_i (here represented by X_i). Because the prevalence of Sp1 is unknown, the frequency of observed presence calculated from PO data is expressed on a relative scale as the probability for $c_{1,i} = 1$ conditioned on $x_{1',jL,i'}$, i.e., Pr ($c_{1',i} = 1 \mid x_{1,jL,i}$).



(a)

1.000	1.000	1.000	1.000	1.000
0.857	0.857	0.857	0.857	0.857
0.714	0.714	0.714	0.714	0.714
0.571	0.571	0.571	0.571	0.571
0.429	0.429	0.429	0.429	0.429
0.286	0.286	0.286	0.286	0.286
0.143	0.143	0.143	0.143	0.143
0.000	0.000	0.000	0.000	0.000

(b)

0.000	0.250	0.500	0.750	1.000
0.000	0.250	0.500	0.750	1.000
0.000	0.250	0.500	0.750	1.000
0.000	0.250	0.500	0.750	1.000
0.000	0.250	0.500	0.750	1.000
0.000	0.250	0.500	0.750	1.000
0.000	0.250	0.500	0.750	1.000
0.000	0.250	0.500	0.750	1.000

(c)

0.963	0.317	0.590	0.571	0.997
0.866	0.374	0.656	0.033	0.831
0.059	0.777	0.924	0.939	0.902
0.325	0.000	0.134	0.654	0.396
1.000	0.145	0.720	0.378	0.200
0.610	1.000	1.000	0.666	0.417
0.896	0.573	0.053	0.531	0.989
0.825	0.479	0.117	0.709	0.305

(d)

0.705	0.487	0.848	0.616	0.551
0.415	0.308	0.867	0.661	0.751
1.000	0.921	0.154	0.853	0.000
0.285	0.328	0.971	0.257	0.770
0.745	0.645	0.540	0.038	0.300
0.089	0.857	0.981	0.835	0.415
0.553	0.708	0.441	0.555	0.726
0.078	0.406	0.203	0.273	0.843

Fig. 5. Example data set 1. (a) Values $x_{1,1\mathrm{L},i}$ for environmental variable (EV) $\boldsymbol{Z}_{1,1}$, which equals the derived variable (DV) $\boldsymbol{X}_{1,1\mathrm{L}}$, in the 40 grid cells in the study area $\boldsymbol{D}_1$. (b) Values $x_{1,2\mathrm{L},i}$ for EV $\boldsymbol{Z}_{1,2}$, which equals DV $\boldsymbol{X}_{1,2\mathrm{L}}$. (c) Values $x_{1,3\mathrm{L},i}$ for EV $\boldsymbol{Z}_{1,3}$, which equals DV $\boldsymbol{X}_{1,3\mathrm{L}}$. (d) Values $x_{1,4\mathrm{L},i}$ for EV $\boldsymbol{Z}_{1,4}$, which equals DV $\boldsymbol{X}_{1,4\mathrm{L}}$. (e) Frequency of observed presence as a measure of the aggregated performance of the modelled target Sp1with respect to $\boldsymbol{X}_{1,1\mathrm{L}}$ and $\boldsymbol{X}_{1,2\mathrm{L}\,.}$ (f) Frequency of observed presence of Sp1 with respect to $\boldsymbol{X}_{1,3\mathrm{L}}$ and $\boldsymbol{X}_{41,4\mathrm{L}}$. Observed presence of Sp1 is indicated by orange-coloured grid cells in (a–d). Red-coloured grid cells in (c) are observed presence cells with $x_{1,3\mathrm{L},i} = 1$.

Example data set 2

The study area for example data set 2 is rasterised into 256 grid cells, arranged in 16 rows × 16 columns (4 × 4 squares, each divided into 4 × 4 grid cells; Fig. 6a). The set of observation units is denoted $\boldsymbol{D}_2 = \{d_{2,1}, ..., d_{2i}, ..., d_{2,256}\}$. A simulated target species 'Sp2' is observed in $n = 48$ (18.75 %) of the $N = 256$ grid cells in $\boldsymbol{D}_2$. No information is available about eventual presence or absence of Sp2 in the remaining $N - n = 208$ uninformed background grid cells. Observed presence is given by the observed presence (OP) vector $\boldsymbol{C}_2 = [c_{2,1}, ..., c_{2,i}, ..., c_{2,256}]^{\mathrm{T}}$. According to the convention for sorting and indexing of grid cells, $i = 1–48$ denote observed presence grid cells, making up the observed presence subset $\boldsymbol{D}_{2+}$, and $i = 49–256$ denote uninformed background grid cells.

The environmental data set $\mathbf{Z}_2$ consists of five explanatory variables, $\boldsymbol{Z}_{2,j}$ ($j = 1, ..., s$; $s = 5$), each recorded for every grid cell in $\boldsymbol{D}_2$; $\boldsymbol{Z}_{2,j} = [z_{2,1j}, ..., z_{2,ij}, ..., z_{2,256j}]^{\mathrm{T}}$. As in example data set 1, $\boldsymbol{Z}_{2,1}$ indexes northing ('Y coordinate') in the rasterised geographical space representation of the study area while $\boldsymbol{Z}_{2,2}$ indexes easting ('X coordinate') in this space (Fig. 6a). $\boldsymbol{Z}_{2,3}$ increases from the 'SW' to the 'NE' corner in each square, thus reflecting finer-scaled variation (Fig. 6a). $\boldsymbol{Z}_{2,4}$ and $\boldsymbol{Z}_{2,5}$ are obtained as vectors of random numbers, drawn from a uniform distribution [0, 1]; $\boldsymbol{Z}_{2,4}$ with the modification that three randomly chosen presence cells were given the maximum value of 1. Explanatory variables $\boldsymbol{Z}_{1,1}$ to $\boldsymbol{Z}_{1,5}$ were ranged to obtain L-type variables $\boldsymbol{X}_{2,1\mathrm{L}}$ ... $\boldsymbol{X}_{2,5\mathrm{L}}$. $\boldsymbol{X}_{2,1\mathrm{L}}$, $\boldsymbol{X}_{2,2\mathrm{L}}$ and $\boldsymbol{X}_{2,3\mathrm{L}}$ are discrete variables with values that make up closed arithmetic sequences starting at 0 and ending at 1, with steps of $\frac{1}{15}$, $\frac{1}{15}$ and $\frac{1}{3}$, respectively (Fig. 6a). $\Pr(x_{2,i1} = \frac{c}{15}) = \Pr(x_{2,i2} = \frac{c}{15}) = 0.0625$ for all integer numbers $c \in [0,15]$ and $\Pr(x_{2,i3} = 0) = \frac{1}{16} = 0.0625$, $\Pr(x_{2,i4} = \frac{1}{3}) = \frac{3}{16} = 0.1875$, $\Pr(x_{2,i3} = \frac{2}{3}) = \frac{5}{16} = 0.3125$ and $\Pr(x_{2,i3} = 1) = \frac{7}{16} = 0.4375$.

The frequency of observed presence of Sp2 with respect to derived variable $\boldsymbol{X}_{2,1\mathrm{L}}$ has a distinct maximum at $z_{2,1} = 0.1875$, decrease rapidly from this maximum towards $x_{2,1\mathrm{L}} = 0$, and levels off gradually towards $x_{2,1\mathrm{L}} = 1$, resulting in a truncated, right-skewed frequency-of-observed-presence curve (Fig. 6b). The frequency-of-observed-presence curve with respect to $\boldsymbol{X}_{2,2\mathrm{L}}$ is irregular, more or less flat-topped for $x_{2,2\mathrm{L}} \le 0.6$, and levels off gradually towards $x_{2,2\mathrm{L}} = 1$ (Fig. 6b). Frequency of observed presence with respect to $\boldsymbol{X}_{2,3\mathrm{L}}$ decreases from 0.4375 for $z_{2,3} = 0$ to 0.1339 for $z_{2,3} = 1$, i.e., from the 'SW' to the 'NE' corner in each square (Fig. 6b). The frequency of observed presence with respect to $\boldsymbol{X}_{2,4\mathrm{L}}$ and $\boldsymbol{X}_{2,5\mathrm{L}}$ varies irregularly (Fig. 6c).

EXPERIMENTS FOR TUNING THE *F*-RATIO TEST

Methods

Two small series of randomisation experiments were performed to evaluate the three alternative values of the parameter η, the effective number of independent observations in MaxEnt models (see the chapter 'The sequential *F*-ratio test'). η is required for determination of the appropriate degrees of freedom to be used in the *F*-ratio test. For each of the two example data sets 1 and 2 ($\boldsymbol{D}_1$ and $\boldsymbol{D}_2$), 100 random derived variables $\boldsymbol{X}_{1,j'}$ and $\boldsymbol{X}_{2,j'}$ ($j' = 1, ..., 100$) of the L type were obtained as sets of 40 and 256 random numbers between 0 and 1. One MaxEnt model without regularisation, $Q_{u,j'}$, where $u = 1, 2$ denotes the example data set, was obtained for each of the 2 × 100 random derived variables by use of customised Excel spreadsheets. For each model, the variation accounted for, $v_{u,j'}$, the fraction of total variation accounted for, $V_{u,j'}$, and the residual variation, $w_{u,j'}$, were obtained by expressions (39), (40) and (41), respectively.

1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.933	0.667	0.667	0.667	1.000	0.667	0.667	0.667	1.000	0.667	0.667	0.667	1.000	0.667	0.667	0.667	1.000
0.867	0.333	0.333	0.667	1.000	0.333	0.333	0.667	1.000	0.333	0.333	0.667	1.000	0.333	0.333	0.667	1.000
0.800	0.000	0.333	0.667	1.000	0.000	0.333	0.667	1.000	0.000	0.333	0.667	1.000	0.000	0.333	0.667	1.000
0.733	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.667	0.667	0.667	0.667	1.000	0.667	0.667	0.667	1.000	0.667	0.667	0.667	1.000	0.667	0.667	0.667	1.000
0.600	0.333	0.333	0.667	1.000	0.333	0.333	0.667	1.000	0.333	0.333	0.667	1.000	0.333	0.333	0.667	1.000
0.533	0.000	0.333	0.667	1.000	0.000	0.333	0.667	1.000	0.000	0.333	0.667	1.000	0.000	0.333	0.667	1.000
0.467	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.400	0.667	0.667	0.667	1.000	0.667	0.667	0.667	1.000	0.667	0.667	0.667	1.000	0.667	0.667	0.667	1.000
0.333	0.333	0.333	0.667	1.000	0.333	0.333	0.667	1.000	0.333	0.333	0.667	1.000	0.333	0.333	0.667	1.000
0.267	0.000	0.333	0.667	1.000	0.000	0.333	0.667	1.000	0.000	0.333	0.667	1.000	0.000	0.333	0.667	1.000
0.200	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.133	0.667	0.667	0.667	1.000	0.667	0.667	0.667	1.000	0.667	0.667	0.667	1.000	0.667	0.667	0.667	1.000
0.067	0.333	0.333	0.667	1.000	0.333	0.333	0.667	1.000	0.333	0.333	0.667	1.000	0.333	0.333	0.667	1.000
0.000	0.000	0.333	0.667	1.000	0.000	0.333	0.667	1.000	0.000	0.333	0.667	1.000	0.000	0.333	0.667	1.000
	0.000	0.067	0.133	0.200	0.267	0.333	0.400	0.467	0.533	0.600	0.667	0.733	0.800	0.867	0.933	1.000

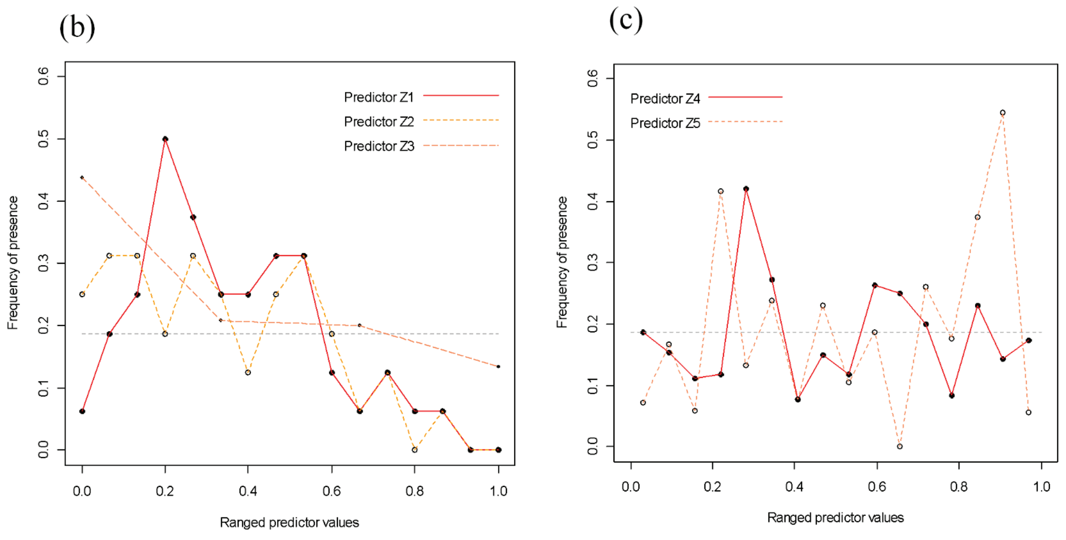


Fig. 6. Example data set 2. (a) The study area $\boldsymbol{D}_2$, rasterised into 16 squares and 256 (16 × 16) grid cells. Values $x_{2,1L,i}$ for environmental variable (EV) $\boldsymbol{Z}_{2,1}$, which equals the derived variable (DV) $\boldsymbol{X}_{2,1L}$, 'northing', are shown along the vertical axis, values $x_{2,2L,i}$ for EV $\boldsymbol{Z}_{2,2}$, which equals DV $\boldsymbol{X}_{2,1L}$, 'easting' , are shown along horizontal axis, and values $x_{2,3L,i}$ for EV $\boldsymbol{Z}_{2,3}$, which equals DV $\boldsymbol{X}_{2,3L}$, originally recorded on an ordinal scale from 1 to 4, are shown in each cell. Observed presence of the target species Sp2 is indicated by orange-coloured cells. (b) Frequency of observed presence of Sp2 with respect to $\boldsymbol{X}_{2,1}$, $\boldsymbol{X}_{2,2}$ and $\boldsymbol{X}_{2,3}$. (c) Frequency of observed presence of Sp2 with respect to $\boldsymbol{X}_{2,4}$ and $\boldsymbol{X}_{2,5}$.

The appropriate value of the parameter η has to satisfy the following three conditions:

1. $\bar{V}_u = \frac{1}{\eta}$ because one random derived variable, on average, 'explains' $\frac{1}{\eta}$ of the total variation (the residual variation after fitting the null model) in a data set.
2. The number of F-ratio tests for comparison of models $Q_{u,j}$ with null models $Q_{u,0}$ for a given value of u that are significant at a given level α is close to 100·α.
3. The two sets of 100 p-values obtained in F-ratio tests for comparison of models $Q_{u,j}$ with null models $Q_{u,0}$ are uniformly distributed over the interval [0,1]. This is motivated by the expectation that, for the correct value of η, the sets of p-values corresponding to the sets of F values are random samples of numbers from [0, 1].

The extent to which alternative values for η satisfied conditions (1) and (2) was evaluated separately for each combination of three η-values and two series of experiments u. Condition (1) was tested by the one-sample t-test (e.g., Crawley 2007): the hypothesis $\bar{V}_u = \frac{1}{\eta}$ was tested against the two-sided alternative hypothesis. The combining probabilities test (Fisher 1954, Sokal & Rohlf 1995) was used to combine the results of the two tests for each value of η ($u = 1, 2$). Condition (2) was evaluated by calculating the p value for each F-ratio test, counting the number of F-ratio tests significant at the α = 0.05 level, and testing this number for deviation from the expected number by an exact binomial test (against the two-sided alternative hypothesis).

For each series of experiments u, the value of η that best satisfied conditions (1) and (2) was tested for consistency with condition (3) by a three-step procedure, using the Kolmogorov-Smirnov (K-S) one-sample test (Sokal & Rohlf 1995): (i) One hundred vectors, each with 100 elements randomly drawn from the uniform distribution [0,1], were obtained. (ii) The vector with 100 p-values obtained in the 100 F-ratio tests of models $Q_{u,j}$ as compared with the null model $Q_{u,0}$, was compared for distributional similarity with each of the 100 random vectors. Since there were 100 random vectors, 100 K-S tests were made. (iii) The number of K-S tests significant at the $\alpha = 0.05$ level was tested for deviation from the expected number of 5 by an exact binomial test (against the one-sided alternative hypothesis, 'greater than').

All statistical analyses other than MaxEnt modelling were performed in R, version 2.11.1 (Anonymous 2010).

Results

The mean fraction of total variation accounted for by the sets of random explanatory variables were $\bar{V}_1 = 0.0327 \pm 0.0042$ (SE of the mean) and $\bar{V}_2 = 0.00495 \pm 0.00076$, respectively. These values corresponded closely to the expected values for $\bar{V}_u$ derived from condition (1) for $\eta = N - n$ ($p = 0.8793$ and 0.8453, cf. Table 6). The test of the hypothesis $\bar{V}_u = \frac{1}{\eta}$ for $\eta = N$ was indicatively significant ($p = 0.0713$) for example data set 1 and not significant ($p = 0.1692$) for example data set 2. The combining probabilities test for $\eta = N$ was indicatively significant ($\chi^2_4 = 8.8551$, $p = 0.0654$).

The number of significant F-ratio tests out of 100 tests for each combination of data set and value for η was close to the expected value of 5 for $\eta = N - n$ (Table 7). For $\eta = 0$, no significant tests were obtained for any of the data sets, while for $\eta = N$ 11 and 7 significant tests were obtained for the two data sets, respectively. Applying the exact binomial test to the results of all 200 F-tests showed that the observed number of significant tests deviated significantly from the expected number, both for $\eta = 0$ and for $\eta = N$ (Table 7). For $\eta = N - n$, the observed number was equal to the expected number, 10.

Frequency distributions for p values obtained in F-ratio tests with $\eta = N - n$ were close

Table 6. One-sample *t*-tests of the null hypothesis $\bar{V}_u = \frac{1}{\eta}$ (against the two-tailed alternative hypothesis) for the three alternative choices of η in the expression for the *F*-ratio. The null hypothesis corresponds to condition (1) for appropriate specification of η. u = example data set, $\bar{V}_u$ = mean fraction of total variation accounted for in MaxEnt models obtained for 100 random derived variables.

Value for η	$u = 1$ ($\bar{V}_u = 0.0327 \pm 0.0042$)			$u = 2$ ($\bar{V}_u = 0.00495 \pm 0.00076$)		
	$\bar{V}_u$ expected	*t* value	*p* value	$\bar{V}_u$ expected	*t* value	*p* value
n	0.1000	−16.026	< 0.0001	0.0208	−20.948	< 0.0001
N	0.0250	1.823	0.0713	0.0039	1.385	0.1692
$N - n$	0.0333	0.152	0.8793	0.0048	0.1957	0.8453

Table 7. Exact binomial tests (against the two-tailed alternative hypothesis) of the null hypothesis that the number of *F*-ratio tests for the given value for the parameter η that are significant at the $\alpha = 0.05$ level equals 5. Tests were made separately for each example data set u and for the two data sets together.

Value for η	$u = 1$		$u = 2$		$u = 1 + 2$	
	# of significant tests	*p* value	# of significant tests	*p* value	# of significant tests	*p* value
n	0	0.0118	0	0.0118	0	< 0.0001
N	11	0.0085	7	0.2559	18	0.0116
$N - n$	6	0.4680	4	0.8719	10	0.8339

to uniform (Fig. 7). For both example data sets, the lowest *p* value obtained for any of the 100 single Kolmogorov-Smirnov tests by which the vector of *p* values from *F*-ratio tests was compared with vectors of 100 random numbers was $p = 0.0541$, giving a *p*-value of 1.0000 for the one-sided exact binomial tests.

$\eta = N - n$ satisfied conditions (1), (2) and (3) for both example data sets, and is therefore most likely to be the appropriate value for η.

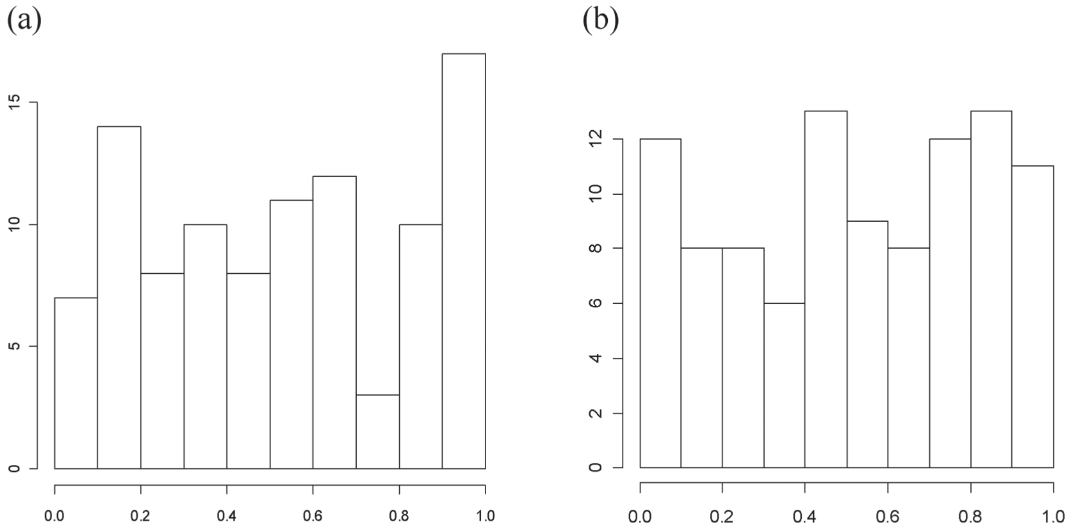


Fig. 7. Frequency distributions (counts) for p values obtained in F-ratio tests (with $\eta = N - n$) for comparison of MaxEnt models $Q_{u,j}$ for 100 random environmental variables with null models $Q_{u,0}$. (a) Example data set 1. (b) Example data set 2.

MAXENT MODELLING OF SIMULATED DATA SETS

Methods

Obtaining derived variables by transformation of explanatory variables

Continuous DVs, i.e., DVs of the linear (L), monotonous (M), and deviation (D) types, were obtained for each of the four explanatory variables (EVs) $\boldsymbol{Z}_{1,j}$ (j = 1, ..., 4) for example data set 1 and for each of the five EVs $\boldsymbol{Z}_{2,j}$ (j = 1, ..., 5) for example data set 2. One DV of each type was derived; the M and D types were represented by the quadratic (Q) and variance (V) valuables, respectively (see Table 2 for transformation functions).

Spline variables, i.e., DVs of the forward hinge (HF), reverse hinge (HR) and threshold (T) types, were constructed only for EVs for which at least one of the MaxEnt models for derived variables of the L, Q, or V types were associated with MaxEnt models that were significantly better than the null model, judged by the F-ratio tests. The significance level α = 0.05 was used as model improvement criterion in all F-ratio tests. Spline DVs were selected by the following procedure: For each of the HF, HR and T variable types, a series of one-variable Maxent models Q_t were obtained without regularisation for knot positions in the open interval (0, 1). The fraction of total variation accounted for, V_t, was calculated for each model in each series by expression (41). A graph of V_t as a function of knot position (referred to as V_t-knot graph) was used to identify eventual local maxima for V_t in the open interval (0,1). HF, HR and T variables were constructed for all knot positions that corresponded to a *distinct* local maximum value for V_t. A local maximum was considered as distinct if it was the maximum of V_t in an interval of breadth 0.2 units, centered on the position of the knot. If no distinct local maximum was found, no spline variable of that type was derived from the explanatory variable in question.

Interaction variables, i.e., DVs of the product (P) and covariance (O) types, were only considered for pairs of EVs that were both represented in the Maxent model (cf. Step 4c in Table 4).

MaxEnt modelling

For each example data set, five MaxEnt models were obtained by standardised procedures as follows:

(1) *By manual forward stepwise selection of DVs and EVs*, following the procedure outlined in Table 4. The following issues were considered in particular detail: (i) comparison of model improvement criteria (Step 1a); (ii) construction of DVs of the spline type (Step 1c); (iii) comparison of frequency-of-observed-presence curves with respect to DVs of different types; and (iv) comparison of predictive performance among DVs of different types (Step 2b).

The sequential *F*-ratio test, given by expression (64), i.e., with $\eta = N - n$, was used as the main model improvement criterion in all comparisons between nested MaxEnt models. Parsimonious sets of DVs for each EV (Step 3) and multi-variable MaxEnt models (Step 4) were obtained by using $\alpha = 0.05$ as criterion for an individually significant contribution to the model. This corresponds to a subset selection regularisation parameter of $\lambda = 3.841$; see the chapter 'Model selection strategies'. In addition, randomisation tests by which each DV was randomised 99999 times, were used to compare all single-variable models with the null model. The difference in AUC_{corr} between nested models ($\Delta\text{AUC}_{\text{corr}}$) was used descriptively for comparison of nested MaxEnt models, i.e., without a statistical test; see the chapter 'The area under the receiver operating curve (AUC)'.

(2,3) *By the 'standard Maxent procedure', using automated selection of DVs and ℓ_1-regularisation, without crossvalidation.* Two models were obtained by the Maxent software, using default options and settings: (2) with L-type variables derived from all EVs; and (3) with all DVs derived from all EVs.

(4,5) *By the 'standard Maxent procedure', using automated selection of DVs and ℓ_1-regularisation, with 5-fold crossvalidation.* Two models parallel to (2) and (3) were obtained by the Maxent software with default options and settings, but with response data (observed presences for the modelled target) divided into five subsets; for $\boldsymbol{D}_1$ with 2 observations each and for $\boldsymbol{D}_2$ with 9, 9, 9, 10 and 10 observations each. Each subset was sequentially left out from the data used to parameterise five MaxEnt submodels; the final MaxEnt model Q was obtained by averaging predictions from the five submodels.

The five MaxEnt models were compared with respect to: (1) the number and identity of DVs included; (2) the fraction of total variation accounted for, V_t; (3) AUC_{corr}, for crossvalidated models 4 and 5 calculated both for training and test data; (4) variable contributions (VC). Four VC measures, one obtained by each of the four procedures outlined in the chapter 'Variable contribution to model', were calculated for each single EV, for sets of DVs derived from one EV, or for single DVs (whichever appropriate): (i) '*Permutation importance*' (Phillips 2011), VC_{PI}, was obtained as the relative reduction in training-data AUC resulting from randomisation of the variable in question, using AUC of the model with this variable as reference. This is the only AUC-based variable contribution measure that was comparable among models with and without crossvalidation. This measure is obtained by a randomisation procedure. (b) '*Percent contribution'* (Phillips 2011), VC_{PC}, obtained by a heuristic method by which the contribution of each variable k to the total variation accounted for by the model is obtained as the sum of changes in variation accounted for (delta log loss) over all steps in the iteration process towards the final model in which the value of the model parameter θ_k was changed; (c) *single-variable*

AUC contribution, $\mathrm{VC}_{\mathrm{AUC}}$, and (d) *single-variable contribution to the total variation accounted for,* $\mathrm{VC}_{\mathrm{FVE}}$, both which is obtained by comparison with the appropriate null model. Single-variable contributions were calculated from separate MaxEnt models for each DV in the full model as the ratio of the DV's contribution and the sum of contributions from all DVs. Contributions of EVs were calculated as the sum of contributions from DVs derived from the EV in question. For models obtained by averaging of several submodels, variable contributions were obtained by averaging all contributions calculated for the five contributing submodels. Plots of 'jackknife variation accounted for' or 'training gain' (Phillips 2011), which are results of resampling procedures provided by the Maxent software, were included for illustration purposes.

MaxEnt modelling was performed partly by use of Maxent software (Phillips et al. 2006, Phillips & Dudík 2008, Phillips 2011), versions 3.3.3e (example data set 1) and 3.3.3k (example data set 2), partly by use of customised Excel spreadsheets. Apart from a couple of noticeable exceptions, results obtained by Maxent software and customised Excel spreadsheets were equal except for rounding errors. Predictions from MaxEnt models are reported in probability-ratio output format, $\dot{q}$, given by expression (84). All statistical analyses other than MaxEnt modelling were performed in R, version 2.11.1 (Anonymous 2010).

Results: example data set 1

MaxEnt reference models

With a total number of $N = 40$ grid cells, of which $n = 10$ are presence cells, log loss values for MaxEnt reference models for example data set 1 were:

The saturated model $Q_{1,S}$ (expression 36): $\ln L_S = \ln n = \ln 10 = 2.3026$
The null model $Q_{1,0}$ (expression 37): $\ln L_0 = \ln N = \ln 40 = 3.6889$

The probability ratio ($\dot{q}_l$) for presence sites as predicted by the saturated MaxEnt model using expression (85) was:

$$\dot{q}_{1,i,S} = \frac{N}{n} = 4,$$

which corresponded to a maximum value of the raw output $q_{1,i,S}$ of $\frac{1}{10}$.

Construction of derived variables of the spline types

$\boldsymbol{Z}_{1,1}$ was the only EV for which DVs of the L, Q or V types resulted in single-variable MaxEnt models significantly better than the null model, judged by the F-ratio test. V_t-knot graphs with at least one distinct local maximum in the interval (0, 1) were obtained for DVs of the HF and T types. The V_t-knot graph for HF-type variables had three local maxima, for knot positions = 0.126, 0.240, and 0.420 (Fig. 8). This graph was discontinuous at knot positions $\frac{3}{7}$, $\frac{4}{7}$, $\frac{5}{7}$ and $\frac{6}{7}$ (Fig. 8), at which values of the transformed DV changed abruptly from a positive value to 0 for all grid cells with $x_{1,1\mathrm{L}}$ values of $\frac{3}{7}$, $\frac{4}{7}$, $\frac{5}{7}$ and $\frac{6}{7}$, respectively. For T-type variables the V_t-knot graph followed a stepwise curve which was discontinuous at knot positions $\frac{1}{7}$, $\frac{2}{7}$, $\frac{3}{7}$, $\frac{4}{7}$, $\frac{5}{7}$ and $\frac{6}{7}$ (figure not shown) and which had a distinct maximum for knot positions in the interval ($\frac{3}{7}$, $\frac{4}{7}$) for which all observed presence grid cells obtained a transformed variable value of 0.

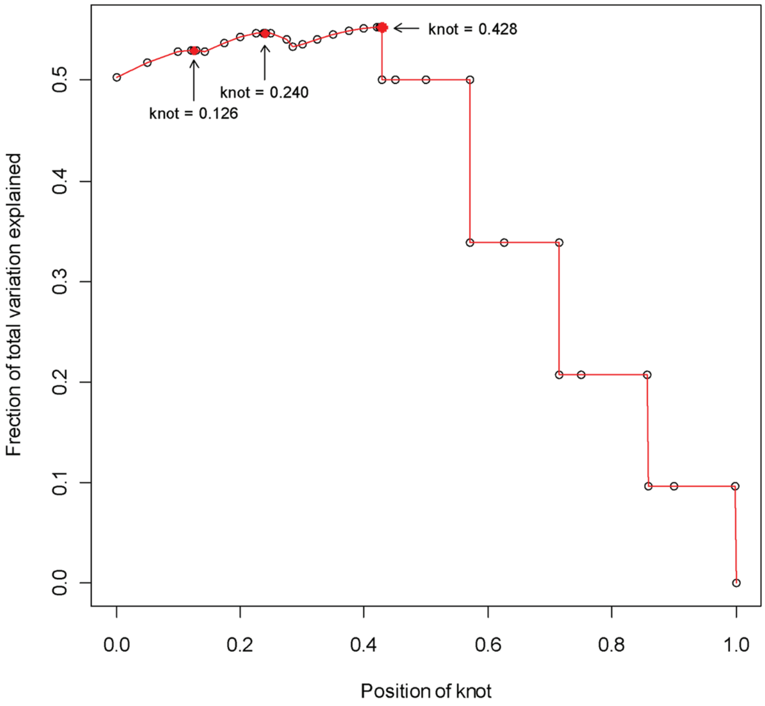


Fig. 8. V_t-knot graph: fraction of total variation explained by single derived variables (DVs) as function of knot position for DVs of the forward-hinge (HF) type, derived from environmental variable $\boldsymbol{Z}_{1,1}$: example data set 1. Distinct local maxima are indicated by red dots.

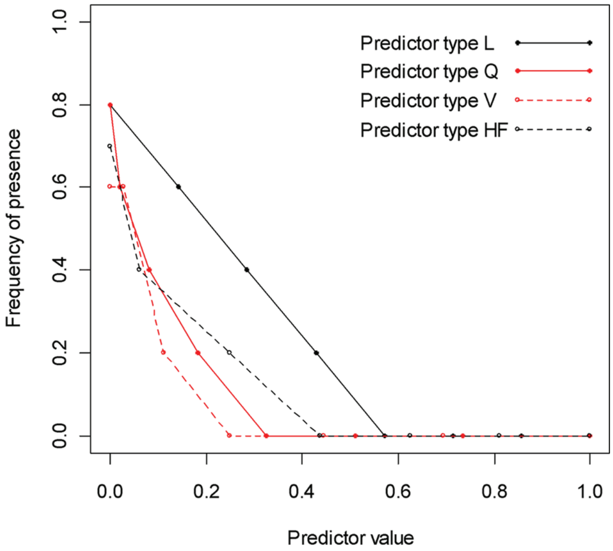


Fig. 9. Frequency-of-observed-presence curves for target species Sp1with respect to derived variables (DVs) of four different types, derived from environmental variable $\boldsymbol{Z}_{1,1}$: example data set 1. The DV of type HF is $x_{1,1\mathrm{HF.240}}$, i.e., the DV with knot at $x_{1,1\mathrm{L}}$ = 0.240.

Frequency-of-observed-presence curves for derived variables of different types

Frequency-of-observed-presence curves for the modelled target Sp1 with respect to the three continuous DVs derived from $\boldsymbol{Z}_{1,1}$ resembled each other (Fig. 9). Compared to the curve for the L variable, the frequency-of-observed-presence curve for the Q variable decreased more rapidly from the maximum at ($x_{1,1Q} = 0$). The curve for the V variable decreased even more rapidly with increasing value of the variable and therefore reached a value of 0 closer to $x_{1,1V} = 0$. The frequency-of-observed-presence curve for the $\mathrm{HF}_{1,\mathrm{HF}.240}$ variable first declined strongly and thereafter (for $x_{1,1\mathrm{HF}.240} > 0.1$) levelled off more gradually. The four curves in Fig. 9 differed with respect to the value of the variable above which no presences were observed, in the order V < Q < HF < L. The T-type variable for a knot in the neighbourhood of $x_{1,1L} = 0.500$ (e.g., 0.450) had frequency of observed presence = 0.5 for $x_{1,1\mathrm{T}.450} = 0$ and frequency of observed presence = 0 for $x_{1,1\mathrm{T}.450} = 1$.

Single-variable models and comparison between model improvement criteria

MaxEnt models for DVs derived from explanatory variable $\boldsymbol{Z}_{1,1}$ accounted for between 49 % and 56 % of the total variation in example data set 1 (Table 8). The highest fraction of total variation accounted for, $V_t = 0.5557$, was obtained for the Q variable. Models for all seven variables derived from $\boldsymbol{Z}_{1,1}$ accounted for significantly more variation than expected by a random DV, judged by both tests (Table 8). The DVs, L, Q and HF.126, obtained the highest $\mathrm{AUC}_{\mathrm{corr}}$ value, $\mathrm{AUC}_{\mathrm{corr}} = 0.916$. Ranked model performance was only weakly correlated (Kendall's $\tau = 0.1588$, $p = 0.6338$, $n = 7$) between V_t and F (which were monotonously related to each other) on one hand and $\mathrm{AUC}_{\mathrm{corr}}$ on the other hand. The highest maximum probability-ratio-output value was obtained for the L variable.

None of the MaxEnt models for continuous DVs derived from EVs $\boldsymbol{Z}_{1,j}$ ($j = 2, ..., 4$) accounted for significantly more variation than expected by a random DV, judged by any test (Table 8). $\mathrm{AUC}_{\mathrm{corr}}$ values for these models ranged from 0.476 for the V variable derived from $\boldsymbol{Z}_{1,3}$ to 0.636 for the L and Q variables derived from $\boldsymbol{Z}_{1,3}$. The corresponding p values in F-ratio and randomisation tests were $0.18 < p < 0.29$.

MaxEnt models parameterised by manual forward stepwise selection

No DV derived from $\boldsymbol{Z}_{1,1}$ accounted for variation that was individually significant when added to the single best DV, $\boldsymbol{X}_{1,1Q}$. All two-variable models had $\mathrm{AUC}_{\mathrm{corr}} = 0.916$. The highest fraction of total variation accounted for by any two-variable model was observed for $\boldsymbol{X}_{1,1\mathrm{T}.450}$ ($V_t = 0.5667$, $\Delta V_t = 0.0110$, $F_{1,26} = 0.662$, $pF = 0.2822$). Based on the model improvement criterion applied, only one parsimonious set of DVs was obtained for example data set 1: the set with the single variable $\boldsymbol{X}_{1,1Q}$ derived from $\boldsymbol{Z}_{1,1}$. Accordingly, the final MaxEnt model obtained by manual forward stepwise selection was the model with $\boldsymbol{X}_{1,1Q}$ as the only DV.

Two-variable MaxEnt models parameterised by adding the variables with highest V_t derived from $\boldsymbol{Z}_{1,2}$ and $\boldsymbol{Z}_{1,3}$ ($\boldsymbol{X}_{1,2L}$ and $\boldsymbol{X}_{1,3Q}$, respectively) to $\boldsymbol{X}_{1,1Q}$ did, however, reveal that DVs can improve multi-variable models significantly even if their individual contribution to explaining variation is not significant according to the single-variable F-ratio test. This was found to be the case for $\boldsymbol{X}_{1,3Q}$ when added to $\boldsymbol{X}_{1,1Q}$ (F-ratio test: $p = 0.0476$; Table 9). The fraction of total variation accounted for by the more complex model with both of $\boldsymbol{X}_{1,1Q}$ and $\boldsymbol{X}_{1,3Q,}$ as DVs was $V_t = 0.6190$, a larger value than the sum of variations accounted for by the models with $\boldsymbol{X}_{1,1Q}$ and $\boldsymbol{X}_{1,3Q}$ as the only derived variables ($V_t = 0.5557$ and $V_t = 0.0479$, respectively, which sum to 0.6036). This shows that fractions of total variation accounted for do not obey the triangular inequality. The variables $\boldsymbol{X}_{1,1Q}$ and $\boldsymbol{X}_{1,3Q}$ were uncorrelated (Kendall's $\tau = 0.0434$, $p = 0.7067$, $n = 40$).

The probability-ratio value $\dot{q}_l = 4$ predicted for presence sites by the saturated MaxEnt

Table 8. Example data set 1: properties of MaxEnt models for single derived variables (DVs), derived from explanatory variables (EVs) $\boldsymbol{Z}_{1,j}$ (j = 1, ..., 4). DVs are coded by type and identity of the EV from which they were derived in accordance with Table 2; for DVs of the spline type the position of the knot is added to the code. V_t = fraction of total variation accounted for; AUC$_{\text{corr}}$ = AUC value, corrected for use with PO data; $F_{\text{df1,df2}}$, pF = F statistic and associated p-value, respectively, for F-ratio tests for comparison of models with the null model; df1 and df2 denote the numbers of degrees of freedom for the numerator and the denominator, respectively, which are 1 and 27 in all tests; $pRand$ = p-value for randomisation tests for comparison of models with the null model (u = 99999 permutations); $\dot{q}_{\text{max}}$ = maximum probability-ratio output value predicted by the model.

EV	DV	V_t	AUC$_{\text{corr}}$	$F_{\text{df1,df2}}$	pF	$pRand$	$\dot{q}_{\text{max}}$
$\boldsymbol{Z}_{1,1}$	$\boldsymbol{X}_{1,1\text{L}}$	0.5030	0.916	27.324	<0.0001	<0.0001	3.947
	$\boldsymbol{X}_{1,1\text{Q}}$	**0.5557**	**0.916**	**33.771**	<0.0001	<0.0001	3.221
	$\boldsymbol{X}_{1,1\text{V}}$	0.5306	0.884	30.521	<0.0001	<0.0001	3.028
	$\boldsymbol{X}_{1,1\text{HF.126}}$	0.5258	0.916	29.939	<0.0001	<0.0001	3.182
	$\boldsymbol{X}_{1,1\text{HF.240}}$	0.5419	0.908	31.943	<0.0001	<0.0001	2.787
	$\boldsymbol{X}_{1,1\text{HF.428}}$	0.5525	0.884	33.331	<0.0001	<0.0001	2.400
	$\boldsymbol{X}_{1,1\text{T.450}}$	0.4943	0.833	26.394	<0.0001	<0.0001	2.000
$\boldsymbol{Z}_{1,2}$	$\boldsymbol{X}_{1,2\text{L}}$	0.0291	0.607	0.810	0.3761	0.2546	1.442
	$\boldsymbol{X}_{1,2\text{Q}}$	0.0286	0.607	0.794	0.3808	0.3161	1.307
	$\boldsymbol{X}_{1,2\text{V}}$	0.0095	0.527	0.259	0.6149	0.5868	1.168
$\boldsymbol{Z}_{1,3}$	$\boldsymbol{X}_{1,3\text{L}}$	0.0422	0.636	1.189	0.2852	0.2305	1.521
	$\boldsymbol{X}_{1,3\text{Q}}$	0.0479	0.636	1.358	0.2551	0.1885	1.678
	$\boldsymbol{X}_{1,3\text{V}}$	0.0058	0.476	0.158	0.6941	0.6967	1.117
$\boldsymbol{Z}_{1,4}$	$\boldsymbol{X}_{1,4\text{L}}$	0.0025	0.533	0.068	0.7963	0.7753	1.139
	$\boldsymbol{X}_{1,4\text{Q}}$	0.0014	0.533	0.039	0.8449	0.8174	1.132
	$\boldsymbol{X}_{1,4\text{V}}$	0.0039	0.540	0.105	0.7484	0.7520	1.106

model was exceeded by both two-variable models (Table 9). Both of these models had larger AUC$_{\text{corr}}$ values than the best one-variable model; the largest value (AUC$_{\text{corr}}$ = 0.937, ΔAUC$_{\text{corr}}$ = 0.021) was observed for the model with $\boldsymbol{X}_{1,1\text{Q}}$ and $\boldsymbol{X}_{1,3\text{Q}}$ (Table 9).

MaxEnt models parameterised by automated variable selection and ℓ_1-regularisation

Maxent auto models without crossvalidation, parameterised by use of L-type variables and by all variables, respectively, contained 2 and 3 derived variables and the fractions of total variation were V_t = 0.5064 and V_t = 0.5482, respectively. The regularised fractions of total variation accounted for (i.e., the fraction of total variation accounted for, calculated from penalised log loss, ln Λ_t, instead of log loss ln L_t) were $\acute{V}_t$ = 0.3619 and $\acute{V}_t$ = 0.5056, respectively. Both of these models had AUC$_{\text{corr}}$ = 0.933. The corresponding models obtained by crossvalidation accounted for less variation that the respective models without crossvalidation despite AUC values were larger; AUC$_{\text{corr}}$ = 0.937 and 0.946, respectively (Table 10).

Table 9. Example data set 1: properties of two-variable MaxEnt models. Derived variables (DVs) are coded by type and identity of the explanatory variable (EV) from which they were derived in accordance with Table 2. V_t = fraction of total variation accounted for by two-variable model; ΔV_t = fraction of total variation accounted for contributed by the added DV; AUC_{corr} = AUC value, corrected for use with presence-only data; $F_{\text{df1,df2}}$, pF = F statistic and associated p-value, respectively, for F-ratio tests for comparison of models with the reference model with $\boldsymbol{X}_{1,1\text{Q}}$ as the only DV; df1 and df2 denote the number of degrees of freedom for the numerator and the denominator, respectively, which are 1 and 26 in both tests; and $\dot{q}_{\max}$ = maximum probability ratio output value predicted by the model.

DV in reference model	DV added	V_t	ΔV_t	AUC_{corr}	$F_{\text{df1,df2}}$	pF	$\dot{q}_{\max}$
$\boldsymbol{X}_{1,1\text{Q}}$	$\boldsymbol{X}_{1,2\text{L}}$	0.5849	0.0291	0.924	1.825	0.1884	4.645
$\boldsymbol{X}_{1,1\text{Q}}$	$\boldsymbol{X}_{1,3\text{Q}}$	0.6190	0.0633	0.937	4.323	0.0476	4.830

Comparison of final MaxEnt models

The final (Man) model obtained by manual forward stepwise selection with $\alpha = 0.05$ in sequential F-ratio tests as model improvement criterion only contained one DV. One additional DV was added to this model if the model improvement criterion was slightly relaxed. The resulting two-variable model is denoted Man+. Models built by the 'standard Maxent procedure' with default options and settings including ℓ_1-regularisation contained 2–3 DVs when built without crossvalidation (Auto|L and Auto|All) and 4–11 DVs when built with crossvalidation (the Auto|L|Xval and Auto|All|Xval models in Table 11).

The fractions of the total variation accounted for by the two Man models, both obtained without ℓ_1-regularisation, and by the four Auto models, all obtained with ℓ_1-regularisation, were of comparable magnitudes; $0.50 < V_t < 0.62$ (Table 11). The highest V_t value was obtained for the two-variable Man+ model. AUC_{corr} values for four of the six models were closely similar (AUC_{corr} = 0.933–0.937) while the 11-variable crossvalidated model parameterised by use of all DVs (Auto|All|Xval) had AUC_{corr} = 0.946 and the manual one-variable model (Man) had AUC_{corr} = 0.916.

Predictions from the six models were strongly correlated; all vectors $\dot{\boldsymbol{Q}}$ of MaxEnt output had pair-wise Kendall's correlation coefficients > 0.88 (Fig. 10). All models distinguished clearly between grid cells with $x_{1,1\text{L}} > 0.5$ for which no presences were recorded and grid cells with $x_{1,1}$ < 0.5 (Fig. 11). The Auto|All|Xval model (Fig. 11d) and the Man+ model (Fig. 11b) stood out from the other models by having vectors of predictions that were very strongly correlated (Fig. 10) and by slightly lower pair-wise correlation coefficients with vectors of predictions from the other models. Predictions from the Man+ model (Fig. 11b), partly also the Auto|All|Xval model (Fig. 11d), varied considerably more among neighbouring grid cells than predictions from the other models, which mainly reflected variation along $\boldsymbol{Z}_{1,1}$. Man+ was the only final model for which predictions for some grid cells exceeded the probability ratio for presence sites in the saturated MaxEnt model given by expression (85), which for example data set 1 was $\dot{q}_i = \frac{N}{n} = 4$ (see Fig. 11b).

Results obtained by the four variable contribution measures were partly inconsistent

Table 10. Example data set 1: properties of final MaxEnt models, parameterised by the 'standard Maxent procedure', i.e., automated selection of derived variables (DVs) and ℓ_1-regularisation with default settings. Model char. = Model characteristics: Auto = model parameterised by the 'standard Maxent procedure' with default options and settings, including ℓ_1-regularisation; L = model parameterised by use of L-type DVs, derived from the four explanatory variables (EVs); All = model parameterised by use of all DVs derived from all EVs by transformations outlined in Table 2 (only one of the HF variables from $\boldsymbol{Z}_{1,1}$, $\boldsymbol{X}_{1,1\mathrm{HF.240}}$ was used); Xval = final model obtained by averaging five models obtained by 5-fold crossvalidation; DV # = number of DVs included in model, the identity of these DVs, coded by type and identity of the EV from which they were derived, is given in a footnote, for models obtained by crossvalidation the number of DVs in single models is given in brackets; $\acute{V}_t$ and V_t = regularised and unregularised fraction of total variation accounted for in a model; $\mathrm{AUC}_{\mathrm{corr}}$ = AUC value, corrected for use with PO data; $F_{\mathrm{df1,df2}}$ and pF are the value of the F statistic and the associated pF value, respectively, for an F-ratio test by which a MaxEnt model is compared with the null model, df1 and df2 denote the number of degrees of freedom for the numerator and the denominator, respectively; $\dot{q}_{\mathrm{max}}$ is the maximum predicted value, as given by the probability-ratio output format.

Model char.	DV #	$\acute{V}_t$	V_t	$\mathrm{AUC}_{\mathrm{corr}}$	df1, df2	$F_{\mathrm{df1,df2}}$	pF	$\dot{q}_{\mathrm{max}}$
Auto|L	2[1]	0.3619	0.5064	0.933	2, 26	13.337	0.0001	3.532
Auto|All	3[2]	0.5056	0.5482	0.933	3, 25	10.111	0.0002	2.388
Auto|L|Xval	4[3] (1–4)	0.3098	0.5342	0.937	–	–	–	3.377
Auto|All|Xval	11[4] (2–5)	0.4610	0.5986	0.946	–	–	–	2.468

DVs: [1]$\boldsymbol{X}_{1,1\mathrm{L}} + \boldsymbol{X}_{1,3\mathrm{L}}$; [2]$\boldsymbol{X}_{1,1\mathrm{Q}} + \boldsymbol{X}_{1,1\mathrm{T.450}} + \boldsymbol{X}_{1,3\mathrm{Q}}$; [3]$\boldsymbol{X}_{1,1\mathrm{L}} + \boldsymbol{X}_{1,2\mathrm{L}} + \boldsymbol{X}_{1,3\mathrm{L}} + \boldsymbol{X}_{1,4\mathrm{L}}$;
[4]$\boldsymbol{X}_{1,1\mathrm{L}} + \boldsymbol{X}_{1,1\mathrm{Q}} + \boldsymbol{X}_{1,1\mathrm{V}} + \boldsymbol{X}_{1,\mathrm{HF.240}} + \boldsymbol{X}_{1,1\mathrm{T.450}} + \boldsymbol{X}_{1,2\mathrm{Q}} + \boldsymbol{X}_{1,3\mathrm{L}} + \boldsymbol{X}_{1,3\mathrm{Q}} + \boldsymbol{X}_{1,3\mathrm{V}} + \boldsymbol{X}_{1,4\mathrm{L}} + \boldsymbol{X}_{1,4\mathrm{V}}$

	Man	Man+	Auto|L	Auto|All	Auto|L|Xval
Man+	0.8825				
Auto|L	0.9479	0.9331			
Auto|All	0.9667	0.9148	0.9806		
Auto|L|Xval	0.9473	0.9016	0.9660	0.9616	
Auto|All|Xval	0.8978	0.8972	0.8936	0.9021	0.9175

Fig. 10. Matrix of Kendall's rank correlation coefficients τ between vectors of predictions from the six Final MaxEnt models for example data set 1 (n = 40). Model characteristics are given in Table 9. All τ correspond to p values < 10^{-10}.

Table 11. Example data set 1: comparison of final MaxEnt models. Model char. = Model characteristics: Man = model parameterised by the manual procedure for forward stepwise selection of derived variables (DVs) and explanatory variables (EVs) outlined in Table 4, using the *F*-ratio test with significance level $\alpha = 0.05$ as model improvement criterion; Man+ = two-variable model obtained from Man by including the marginally significant DV $\boldsymbol{X}_{1,3Q}$ in addition to $\boldsymbol{X}_{1,1Q}$; Auto = model parameterised by the 'standard Maxent procedure' with default options and settings, including ℓ_1-regularisation; L = model parameterised by use of L-type DVs derived from the four EVs; All = model parameterised by use of all DVs derived from all EVs by transformations outlined in Table 2; (only one of the HF-type variables from $\boldsymbol{Z}_{1,1}$, $\boldsymbol{X}_{1,1HF.240}$ was used); Xval = final model obtained by averaging five models obtained by 5-fold crossvalidation]; DV # = number of DVs included in model, the identity of these DVs, coded by type and identity of the EV from which they were derived, is given in a footnote, for models obtained by crossvalidation the number of DVs in single models is given in brackets; V_t = (unregularised) fraction of total variation accounted for by a model; AUC_{corr} = AUC value, corrected for use with PO data; VC_{PI}, VC_{PC}, VC_{AUC} and VC_{FVA} = variable contributions calculated for each EV by four different measures (see text for explanation), expressed as fractions of the sum of contributions by all EVs.

Model char.	DV #	V_t	AUC_{corr}	EV	VC_{PI}	VC_{PC}	VC_{AUC}	VC_{FVA}
Man	1[1]	0.5557	0.916	$\boldsymbol{Z}_{1,1}$	1.000	1.000	1.000	1.000
				$\boldsymbol{Z}_{1,2}$	–	–	–	–
				$\boldsymbol{Z}_{1,3}$	–	–	–	–
				$\boldsymbol{Z}_{1,4}$	–	–	–	–
Man+	2[2]	0.6190	0.937	$\boldsymbol{Z}_{1,1}$	1.000	0.902	0.746	0.921
				$\boldsymbol{Z}_{1,2}$	–	–	–	–
				$\boldsymbol{Z}_{1,3}$	0.000	0.098	0.254	0.079
				$\boldsymbol{Z}_{1,4}$	–	–	–	–
Auto|L	2[3]	0.5064	0.933	$\boldsymbol{Z}_{1,1}$	0.317	0.987	0.754	0.879
				$\boldsymbol{Z}_{1,2}$	0.683	0.012	–	0.045
				$\boldsymbol{Z}_{1,3}$	0.000	0.001	0.246	0.076
				$\boldsymbol{Z}_{1,4}$	–	–	–	–
Auto|All	3[4]	0.5482	0.933	$\boldsymbol{Z}_{1,1}$	0.936	1.000	0.846	0.956
				$\boldsymbol{Z}_{1,2}$	–	–	–	–
				$\boldsymbol{Z}_{1,3}$	0.064	0.000	0.154	0.044
				$\boldsymbol{Z}_{1,4}$	–	–	–	–
Auto|L|Xval	4[5] (1–4)	0.5342	0.937	$\boldsymbol{Z}_{1,1}$	0.318	0.960	0.601	0.931
				$\boldsymbol{Z}_{1,2}$	0.214	0.003	0.155	0.022
				$\boldsymbol{Z}_{1,3}$	0.282	0.013	0.197	0.047
				$\boldsymbol{Z}_{1,4}$	0.186	0.025	0.048	0.001
Auto|All|Xval	11[6] (2–5)	0.5986	0.946	$\boldsymbol{Z}_{1,1}$	0.878	0.982	0.906	0.978
				$\boldsymbol{Z}_{1,2}$	0.030	0.011	0.024	0.005
				$\boldsymbol{Z}_{1,3}$	0.055	0.006	0.055	0.017
				$\boldsymbol{Z}_{1,4}$	0.037	0.001	0.015	0.001

DVs: [1]$\boldsymbol{X}_{1,1Q}$; [2]$\boldsymbol{X}_{1,1Q} + \boldsymbol{X}_{1,3Q}$; [3]$\boldsymbol{X}_{1,1L} + \boldsymbol{X}_{1,3L}$; [4]$\boldsymbol{X}_{1,1Q} + \boldsymbol{X}_{1,1T.450} + \boldsymbol{X}_{1,3Q}$; [5]$\boldsymbol{X}_{1,1L} + \boldsymbol{X}_{1,2L} + \boldsymbol{X}_{1,3L} + \boldsymbol{X}_{1,4L}$; [6]$\boldsymbol{X}_{1,1L} + \boldsymbol{X}_{1,1Q} + \boldsymbol{X}_{1,1V} + \boldsymbol{X}_{1,HF.240} + \boldsymbol{X}_{1,1T.450} + \boldsymbol{X}_{1,2Q} + \boldsymbol{X}_{1,3L} + \boldsymbol{X}_{1,3Q} + \boldsymbol{X}_{1,3V} + \boldsymbol{X}_{1,4L} + \boldsymbol{X}_{1,4V}$

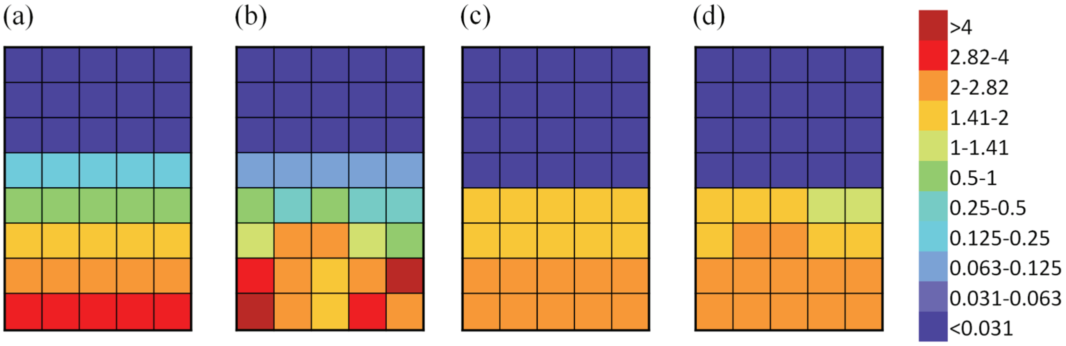


Fig. 11. Map representation of predictions for the modelled target Sp1 in example data set 1, given in probability-ratio output format $\dot{q}$. (a) The Man model, parameterized by the manual procedure for forward stepwise selection of derived variables (DVs) and explanatory variables (EVs) outlined in Table 4, using the F-ratio test with significance level $\alpha = 0.05$ as model improvement criterion. (b) The Man+ model, a two-variable model obtained from the Man model by including the marginally significant DV $\boldsymbol{X}_{1,3Q}$ in addition to $\boldsymbol{X}_{1,1Q}$. (c) The Auto|All model, parameterized by the 'standard Maxent procedure' with default options and settings, including ℓ_1-regularisation, by use of all DVs derived from all EVs by transformations outlined in Table 2. (d) The Auto|All|Xval model, which is similar to the Auto|All model except for being built with 5-fold crossvalidation.

(Table 11). Permutation importance (VC_{PI}) deviated most strongly from the others, by its estimated contribution of only ca. 0.3 from $\boldsymbol{Z}_{1,1}$ to the Auto|L and Auto|L|Xval models. In contrast, the contribution from $\boldsymbol{Z}_{1,1}$ was larger than 0.8 according to the two measures based upon fraction of total variation accounted for (VC_{FVE} and VC_{PC}). According to the AUC-based measure (VC_{AUC}), contributions from $\boldsymbol{Z}_{1,1}$ were intermediate between these extremes ($VC_{AUC} > 0.6$). This was in accordance with the *relatively* much higher ΔAUC_{corr} (compared with the null model AUC_{corr} of 0.5) than V_t values for DVs derived from EVs $\boldsymbol{Z}_{1,2}$, $\boldsymbol{Z}_{1,3}$ and $\boldsymbol{Z}_{1,4}$ than from $\boldsymbol{Z}_{1,1}$ (Table 8).

Results: example data set 2

MaxEnt reference models

With a total number of $N = 256$ grid cells, of which $n = 48$ are presence cells, log loss values for MaxEnt reference models for example data set 2 were:

The saturated model $Q_{2,S}$ (expression 36): $\ln L_S = \ln n = \ln 48 = 3.8712$
The null model $Q_{2,0}$ (expression 37): $\ln L_0 = \ln N = \ln 256 = 5.5452$

The probability ratio ($\dot{q}_l$) for presence sites as predicted by the saturated MaxEnt model using expression (85) was:

$$\dot{q}_{2,i,S} = \frac{N}{n} = 5.333,$$

which corresponded to a maximum value of the raw output $q_{2,i,S}$ of $\frac{1}{48}$.

Construction of derived variables of the spline types

Single-variable MaxEnt models that were significantly better than the null model, judged by the F-ratio test, were obtained for continuous DVs, i.e., DVs of the L, Q or V types, derived from the three EVs $\boldsymbol{Z}_{2,1}$, $\boldsymbol{Z}_{2,2}$ and $\boldsymbol{Z}_{2,3}$. V_t-knot graphs with one distinct local maximum in the interval (0, 1) were obtained for HF, HR and T variables derived from $\boldsymbol{Z}_{2,1}$ (Fig. 12a) and for HF and T variables derived from $\boldsymbol{Z}_{2,2}$ (Fig. 12b). Additional, indistinct local minima could be observed on some V_t-knot graphs. V_t-knot graphs for HF-type variables derived both from $\boldsymbol{Z}_{2,1}$ and $\boldsymbol{Z}_{2,2}$ were discontinuous at knot position = $\frac{15}{16}$ (Fig. 12). For T-type variables, the V_t-knot graph followed a stepwise curve which was discontinuous at knot positions $\frac{u}{16}$ where u = 1, ..., 15 (not shown in Fig. 12). The T-type variable derived from $\boldsymbol{Z}_{2,1}$ had two distinct local maxima, of which the lesser (for knot position ≈ 0.1) coincided with the (global) maximum for the HR-type variable and the greater (for knot position ≈ 0.650) coincided with the maximum for the HF-type variable. The T-type variable derived from $\boldsymbol{Z}_{2,2}$ had one distinct maximum only, for a position of the knot close to the position of the maximum for the HF-type variable. No distinct maximum within (0,1) was found for the HR-type variable.

One spline variable (of the HR type) derived from explanatory variable $\boldsymbol{Z}_{2,3}$ had a distinct local maximum in (0, 1).

Frequency-of-observed-presence curves for derived variables of different types

Frequency-of-observed-presence curves for the modelled target Sp2 with respect to the three continuous variables derived from $\boldsymbol{Z}_{2,1}$ resembled each other (Fig. 13a); all had one distinct mode. This mode was displaced towards lower values for the derived variable from the L via the Q to the V variable. Frequency-of-observed-presence curves with respect to DVs of the HR and HF types with knots at $x_{2,1\mathrm{L}}$ = 0.150 and $x_{2,1\mathrm{L}}$ = 0.600, respectively, were also closely similar; frequency maxima (> 0.2) were found for the value 0 of the respective DVs from which the frequency of observed presence gradually decreased. For values of the derived variables = 1, frequencies < 0.1 were observed. The $\boldsymbol{X}_{2,1\mathrm{T}.650}$ variable maximised the difference in frequency of observed presence between grid cells with $x_{2,1\mathrm{T}.650}$ = 1, $\bar{x}_{2,2\mathrm{T}}{}^{*}$ = 0.0729 and grid cells with $x_{2,1\mathrm{T}.650}$ = 0; 0.2562.

Frequency-of-observed-presence curves with respect to the three continuous variables derived from $\boldsymbol{Z}_{2,2}$ also resembled each other closely (Fig. 13b); frequencies of observed presence around the maximum value of ca. 0.25 were observed for low values of the DVs while the frequencies gradually decreased with increasing DV values. The broadest interval with frequency of observed presence around the maximum was observed for the L variable (0–0.65) while the narrowest interval was observed for the V variable (0–0.30). The frequency-of-observed-presence curve with respect to $\boldsymbol{X}_{2,2\mathrm{HF}.612}$ levelled off more gradually than the corresponding curves for continuous DVs (Fig. 13b). The $\boldsymbol{X}_{2,2\mathrm{T}.700}$ variable maximised the difference in frequency of observed presence between grid cells with $x_{2,2\mathrm{T}.700}$ = 1, $\bar{x}_{2,2\mathrm{T}}{}^{*}$ = 0.05, and grid cells with $x_{2,2\mathrm{T}.700}$ = 0; 0.25.

Single-variable models and comparison between model improvement criteria

MaxEnt models for DVs derived from EV $\boldsymbol{Z}_{2,1}$ accounted for between 2 % and 13 % of the total variation in example data set 2 (Table 12). The strong gradient in fraction of total variation accounted for from $V_{2,1\mathrm{L}}$ = 0.0216 via $V_{2,1\mathrm{Q}}$ = 0.0511 to $V_{2,1\mathrm{V}}$ = 0.1262 coincided with the displacement of frequency-of-observed-presence maxima from $x_{2,1\mathrm{L}}$ = 0.267 via $x_{2,1\mathrm{Q}}$ = 0.040 to $x_{2,1\mathrm{V}}$ = 0.026 (Fig. 13a). The peak of the frequency-of-observed-presence curves was less sharp (and the curve closer to monotonous) for $\boldsymbol{X}_{2,1\mathrm{V}}$ due to higher frequency of observed presence for $x_{2,1\mathrm{V}}$ = 0 than for $x_{2,1\mathrm{L}}$ = $x_{2,1\mathrm{Q}}$ = 0. $\boldsymbol{X}_{2,1\mathrm{V}}$ performed the best among DVs derived from $\boldsymbol{Z}_{2,1}$, judged by all

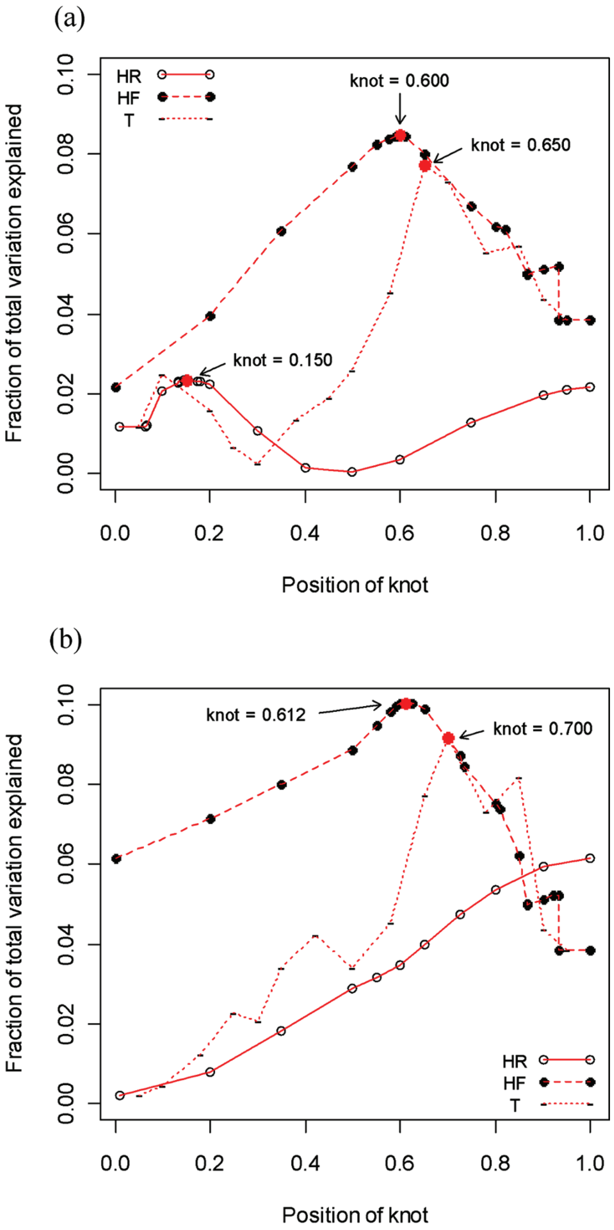


Fig. 12. V_t-knot graphs: fraction of total variation explained by single derived variables (DVs) as function of knot position for DVs of the spline main type, i.e., DVs of the reverse hinge (HR), forward hinge (HF) and threshold (T) types, derived from environmental variables (EVs): example data set 2. (a) DVs derived from EV $\boldsymbol{Z}_{2,1}$. (b) DVs derived from EV $\boldsymbol{Z}_{2,2}$. Distinct local maxima are indicated by red dots.

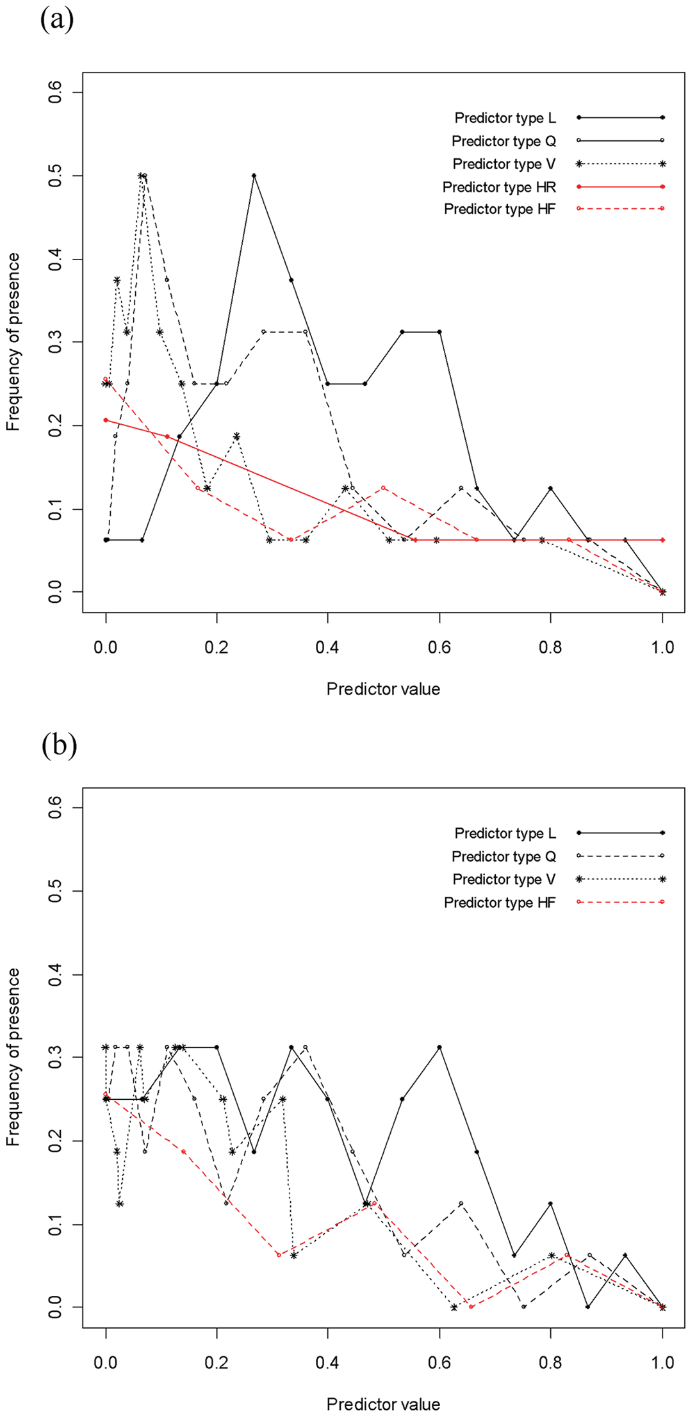


Fig. 13. Frequency-of-observed-presence curves for target species Sp2 with respect to derived variables (DVs) of different types derived from environmental variables (EVs): example data set 2. (a) EV $\boldsymbol{Z}_{2,1}$. DVs of types HR and HF are $x_{2,1\mathrm{HR}.150}$ and $x_{2,1\mathrm{HF}.600}$, respectively. (b) EV $\boldsymbol{Z}_{2,2}$. The DV of type HF is $x_{2,2\mathrm{HF}.612}$.

Table 12. Example data set 2: properties of Maxent models for single derived variables (DVs), derived from explanatory variables (EVs) $\boldsymbol{Z}_{2,j}$ (j = 1, ..., 5). DVs are coded by type and identity of the EV from which they were derived in accordance with Table 2; for DVs of the spline type the position of the knot is added to the code. V_t = fraction of total variation accounted for; AUC$_{\text{corr}}$ = AUC value, corrected for use with PO data; $F_{\text{df1,df2}}$, pF = F statistic and associated p-value, respectively, for F-ratio tests for comparison of models with the null model; df1 and df2 denote the number of degrees of freedom for the numerator and the denominator, respectively, which are 1 and 205 in all tests; *pRand* = p-value for randomisation tests for comparison of models with the null model (u = 99999 permutations); $\dot{q}_{\max}$ = maximum probability-ratio output value predicted by the model.

EV	IV	V_t	AUC$_{\text{corr}}$	$F_{\text{df1,df2}}$	pF	*pRand*	$\dot{q}_{\max}$
$\boldsymbol{Z}_{2,1}$	$\boldsymbol{X}_{2,1\text{L}}$	0.0216	0.596	4.528	0.0345	0.0370	1.500
	$\boldsymbol{X}_{2,1\text{Q}}$	0.0511	0.596	11.047	0.0011	0.0015	1.510
	$\boldsymbol{X}_{2,1\text{V}}$	**0.1262**	**0.708**	**29.598**	<0.0001	<0.0001	1.880
	$\boldsymbol{X}_{2,1\text{HR.150}}$	0.0225	0.554	4.723	0.0309	0.0409	1.100
	$\boldsymbol{X}_{2,1\text{HF.600}}$	0.0836	0.650	18.692	<0.0001	<0.0001	1.340
	$\boldsymbol{X}_{2,1\text{T.650}}$	0.0764	0.642	16.955	<0.0001	<0.0001	1.365
$\boldsymbol{Z}_{2,2}$	$\boldsymbol{X}_{2,2\text{L}}$	0.0614	0.660	13.410	0.0003	0.0005	1.924
	$\boldsymbol{X}_{2,2\text{Q}}$	0.0811	0.660	18.104	<0.0001	<0.0001	1.667
	$\boldsymbol{X}_{2,2\text{V}}$	0.0774	0.645	17.195	<0.0001	0.0001	1.590
	$\boldsymbol{X}_{2,2\text{HF.612}}$	**0.0989**	0.655	**22.488**	<0.0001	<0.0001	1.364
	$\boldsymbol{X}_{2,2\text{T.700}}$	0.0908	0.642	20.470	<0.0001	<0.0001	1.365
$\boldsymbol{Z}_{2,3}$	$\boldsymbol{X}_{2,3\text{L}}$	**0.0312**	0.602	**6.595**	0.0109	0.0129	1.911
	$\boldsymbol{X}_{2,3\text{Q}}$	0.0260	0.602	5.475	0.0203	0.0230	1.523
	$\boldsymbol{X}_{2,3\text{V}}$	0.0048	0.504	0.982	0.3229	0.3139	1.388
	$\boldsymbol{X}_{2,3\text{HR.445}}$	0.0287	0.573	6.049	0.0147	0.0102	2.321
$\boldsymbol{Z}_{2,4}$	$\boldsymbol{X}_{2,4\text{L}}$	0.0004	0.511	0.088	0.7670	0.7557	1.070
	$\boldsymbol{X}_{2,4\text{Q}}$	0.0007	0.511	0.140	0.7087	0.7061	1.057
	$\boldsymbol{X}_{2,4\text{V}}$	0.0005	0.511	0.100	0.7522	0.7458	1.044
$\boldsymbol{Z}_{2,5}$	$\boldsymbol{X}_{2,5\text{L}}$	0.0090	0.566	1.853	0.1749	0.1780	1.329
	$\boldsymbol{X}_{2,5\text{Q}}$	0.0087	0.566	1.791	0.1823	0.1808	1.425
	$\boldsymbol{X}_{2,5\text{V}}$	0.0006	0.485	0.130	0.7188	0.7119	1.049

performance statistics ($V_{2,1\text{V}}$ = 0.1262; AUC$_{\text{corr}}$ = 0.708 which was 0.058 AUC$_{\text{corr}}$ units higher than the value observed for the second best DV, $\boldsymbol{X}_{2,1\text{HF.600}}$; Table 12). Models for all six DVs derived from $\boldsymbol{Z}_{2,1}$ did, however, explain significantly more variation than expected by a random DV, judged by both tests (although tests for $\boldsymbol{X}_{2,1\text{L}}$ and $\boldsymbol{X}_{2,1\text{HR.150}}$ were only marginally significant; Table 12). The broad patterns of ranked model performance given by V_t and F on one hand and by AUC on the other hand were quite similar although $\boldsymbol{X}_{2,1\text{L}}$ and $\boldsymbol{X}_{2,1\text{HR.150}}$, which had similar V_t values, differed with respect to AUC$_{\text{corr}}$ by 0.042 units, and the 2.5-fold increase in variation accounted for from

$\boldsymbol{X}_{2,1\mathrm{L}}$ to $\boldsymbol{X}_{2,1\mathrm{Q}}$ was not reflected in AUC$_{\mathrm{corr}}$.

The variation accounted for by the five DVs derived from $\boldsymbol{Z}_{2,2}$ varied between narrow limits (from $V_{2,2\mathrm{L}} = 0.0614$ to $V_{2,2\mathrm{HF.612}} = 0.0989$), as expected from their highly similar frequency-of-observed-presence curves (Fig. 13b). All DVs derived from $\boldsymbol{Z}_{2,2}$ accounted for significantly more variation than expected by a random DV, judged both by the *F*-ratio test and the randomisation test ($p < 0.001$; Table 12). No correspondence was observed between V_t and AUC.

The linear variable $\boldsymbol{X}_{2,3\mathrm{L}}$ accounted for most variation ($V_{2,3\mathrm{L}} = 0.0312$) among DVs derived from $\boldsymbol{Z}_{2,3}$, and was judged best by the *F*-ratio test and by AUC$_{\mathrm{corr}}$ and second best by the randomisation test (Table 12). While $\boldsymbol{X}_{2,3\mathrm{L}}$, $\boldsymbol{X}_{2,3\mathrm{Q}}$ and $\mathbf{X}_{2,3\mathrm{HR.445}}$ were similar with respect to all performance measures, $\boldsymbol{X}_{2,3\mathrm{V}}$ accounted for a negligible fraction of variation, not significantly more than expected by a random DV (Table 12).

None of the MaxEnt models for continuous DVs derived from EVs $\boldsymbol{Z}_{2,4}$ and $\boldsymbol{Z}_{2,5}$ accounted for significantly more variation than expected by a random DV, judged by any test (Table 12). The AUC of these models ranged from 0.485 for $\boldsymbol{X}_{2,5\mathrm{V}}$ to 0.566 for $\boldsymbol{X}_{2,5\mathrm{L}}$ and $\boldsymbol{X}_{2,5\mathrm{Q}}$.

For the entire set of 21 single-variable models for example data set 2, *p* values of the *F*-ratio and randomisation tests were very closely similar (Table 12), both in terms of numerical values (Pearson's product-moment correlation coefficient $r = 0.99992$, $p < 10^{-16}$, $n = 14$; models with at least one *p* value < 0.0001 not included) and rank order (Kendall's rank correlation coefficient $\tau = 0.9560$, $p < 10^{-8}$, $n = 14$). The fraction of total variation accounted for (V_t) and AUC$_{\mathrm{corr}}$ were nonlinearly related (Fig. 14); AUC$_{\mathrm{corr}}$ increased by ca. 0.1 units, from 0.5 to 0.6 and from 0.6 to 0.7, respectively, in response to increase of V_t from 0 to 0.03 and from 0.03 to ca. 0.12, respectively. Nevertheless, V_t and AUC$_{\mathrm{corr}}$ were significantly correlated (Kendall's $\tau = 0.7671$, $p = 2 \cdot 10^{-6}$, $n = 21$).

MaxEnt models parameterised by manual forward stepwise selection

DVs with individually significant contributions to variation accounted for were derived from the three EVs $\boldsymbol{Z}_{2,1}$, $\boldsymbol{Z}_{2,2}$, $\boldsymbol{Z}_{2,3}$. No DV gave an individually significant contribution to variation accounted for when added to the single best DV derived from the same EV (all $pF > 0.4$, Δ AUC$_{\mathrm{corr}} < 0.008$; Table 13). The parsimonious sets of DVs derived from $\boldsymbol{Z}_{2,1}$, $\boldsymbol{Z}_{2,2}$ and $\boldsymbol{Z}_{2,3}$ therefore consisted of one DV each.

Two-variable MaxEnt models parameterised by adding one parsimonious set of DVs derived from $\boldsymbol{Z}_{2,2}$ or $\boldsymbol{Z}_{2,3}$, consisting of one derived variable each, to the best one-variable model, i.e., the model with $\boldsymbol{X}_{2,1\mathrm{V}}$ as the only derived variable, both accounted for significantly more variation than expected of models obtained by adding a random DV, judged by the *F*-ratio test (Table 14). EVs $\boldsymbol{Z}_{2,1}$ and $\boldsymbol{Z}_{2,2}$, from which variables $\boldsymbol{X}_{2,1\mathrm{V}}$ and $\boldsymbol{X}_{2,2\mathrm{HF.612}}$ were derived, were orthogonal ($r = 0$). The increase in fraction of total variation accounted for resulting from adding $\boldsymbol{X}_{2,2\mathrm{HF.612}}$ to $\boldsymbol{X}_{2,1\mathrm{V}}$, $\Delta V_t = 0.0989$, was therefore equal to the variation accounted for by the single-variable model with $\boldsymbol{X}_{2,2\mathrm{HF.612}}$ (compare Tables 12 and 14). EVs $\boldsymbol{Z}_{2,1}$ and $\boldsymbol{Z}_{2,3}$ were weakly correlated ($r = 0.1462$, $p = 0.0192$, $n = 256$), and the increase in ΔV_t resulting from adding $\boldsymbol{X}_{2,3\mathrm{L}}$ to $\boldsymbol{X}_{2,1\mathrm{V}}$, $\Delta V_t = 0.0293$, was slightly lower than the variation accounted for by the single-variable model with $\boldsymbol{X}_{2,3\mathrm{L}}$ ($V_{2,3\mathrm{L}} = 0.0312$). The increase in AUC$_{\mathrm{corr}}$ resulting from adding $\boldsymbol{X}_{2,2\mathrm{HF.612}}$ and $\boldsymbol{X}_{2,3\mathrm{L}}$, respectively, to the model, were 0.080, and 0.039, respectively (Table 14). P- and O-type variables obtained from $\boldsymbol{Z}_{2,1}$ and $\boldsymbol{Z}_{2,2}$ did not significantly improve the best two-variable model (Table 14).

The final MaxEnt model obtained by manual forward stepwise selection was the model with the three DVs $\boldsymbol{X}_{2,1\mathrm{V}}$, $\boldsymbol{X}_{2,2\mathrm{HF.612}}$ and $\boldsymbol{X}_{2,3\mathrm{L}}$ (Table 14). Adding $\boldsymbol{X}_{2,3\mathrm{L}}$ to the best two-variable model increased the fraction of total variation accounted for by $\Delta V_t = 0.0185$, AUC$_{\mathrm{corr}}$ increased by 0.020 units, and the *F*-ratio test was significant at the $p < 0.05$ level (Table 14). P- and O-type variables obtained from $\boldsymbol{Z}_{2,1}$ and $\boldsymbol{Z}_{2,3}$, or from $\boldsymbol{Z}_{2,2}$ and $\boldsymbol{Z}_{2,3}$, were unlikely to improve the best three-variable model significantly, and were not tested.

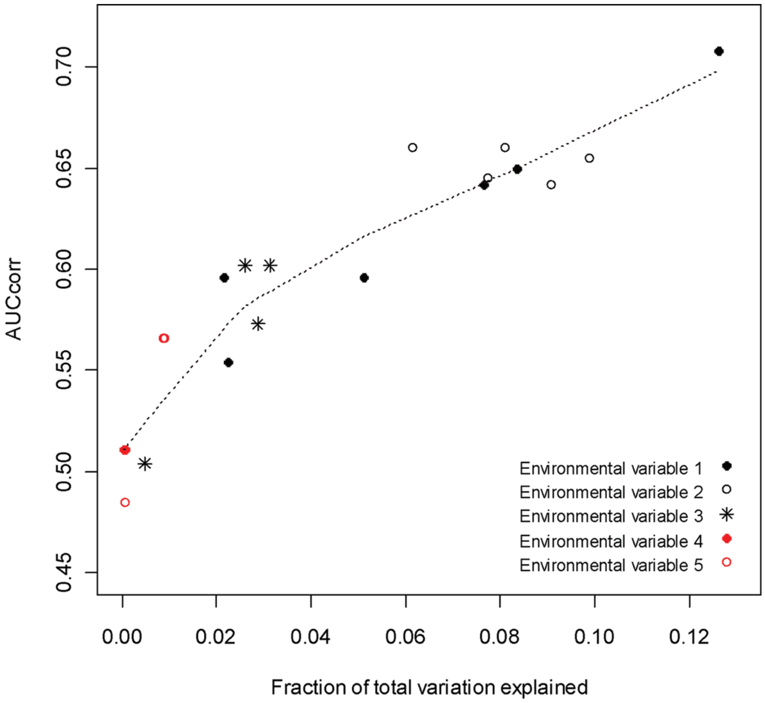


Fig. 14. Example data set 2: relationship between AUC$_{\text{corr}}$, i.e., AUC corrected for use with PO data, and fraction of total variation explained, V_t, for single-variable MaxEnt models for the 21 derived variables of different types derived from the five environmental variables. The trendline is a lowess smoother.

The maximum probability-ratio ($\dot{q}_l$) output value predicted by the models increased with increasing model complexity, but remained below the value predicted for observed presence sites by the saturated MaxEnt model (5.333) for all models (Table 14).

MaxEnt models parameterised by automated variable selection and ℓ_1-regularisation

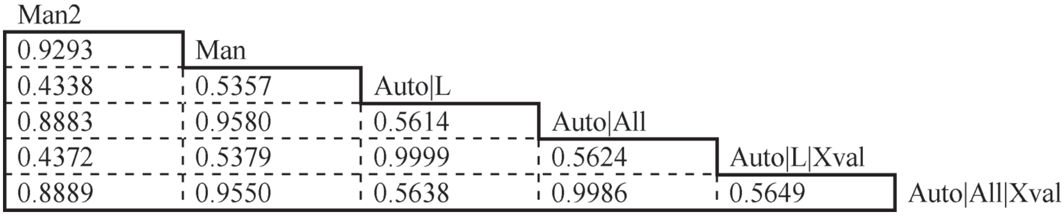
Maxent auto models parameterised by use of L-type variables, without and with crossvalidation, respectively, were closely similar with respect to fractions of the total variation accounted for (V_t = 0.268–0.275) and AUC (AUC$_{\text{corr}}$ = 0.720–0.725). Furthermore, vectors of predictions from these two models were closely similar (Kendall's $\tau > 0.999$, $n = 256$; Fig. 15). Maxent auto models parameterised by use of all 21 DVs derived from the five EVs, without and with crossvalidation, respectively, were also similar with respect to V_t (0.26–0.30), AUC$_{\text{corr}}$ (0.81–0.83) and vectors of model predictions (Kendall's $\tau > 0.998$, $n = 256$; Fig. 15) (Table 15). The model obtained without crossvalidation had nonzero coefficients for 11 DVs while all 21 DVs had nonzero coefficients in a least one of the five single models that contributed to the crossvalidated model.

Comparison of final MaxEnt models

Because models with and without crossvalidation in the two pairs (L and All models) were closely similar (Fig. 15), only models without crossvalidation (Auto|L and Auto|All) were compared with models obtained by manual forward stepwise selection. Also the model Man2, which is similar

Table 13. Example data set 2: selection of parsimonious sets of derived variables (DVs) for each explanatory variable (EV): properties of two-variable MaxEnt models. DVs are coded by type and identity of the EV from which they were derived in accordance with Table 2. V_t = fraction of total variation accounted for by the model; ΔV_t = fraction of total variation accounted for by the added DV; AUC$_{\mathrm{corr}}$ = AUC value, corrected for use with PO data; Δ AUC$_{\mathrm{corr}}$ = difference in AUC$_{\mathrm{corr}}$ between the two-variable model and the one-variable model used as reference; $F_{\mathrm{df1,df2}}$, pF = F statistic and associated p-value, respectively, for F-ratio tests for comparison of models with a reference model with only one DV; df1 and df2 denote the number of degrees of freedom for the numerator and the denominator, respectively, which are 1 and 204 in all tests; $\dot{q}_{\mathrm{max}}$ = maximum probability-ratio output value predicted by the model. Properties of the best one-variable reference model are shown on gray background.

DV in ref. model	DVadded	V_t	ΔV_t	AUC$_{\mathrm{corr}}$	Δ AUC$_{\mathrm{corr}}$	$F_{\mathrm{df1,df2}}$	pF	$\dot{q}_{\mathrm{max}}$
$\boldsymbol{X}_{2,1\mathrm{V}}$	none	*0.1262*	–	*0.708*	–	–	–	*1.880*
$\boldsymbol{X}_{2,1\mathrm{V}}$	$\boldsymbol{X}_{2,1\mathrm{HF.600}}$	0.1262	0	0.708	0	0.001	0.9748	1.878
$\boldsymbol{X}_{2,1\mathrm{V}}$	$\boldsymbol{X}_{2,1\mathrm{T.650}}$	0.1291	0.0030	0.708	0	0.693	0.4061	1.874
$\boldsymbol{X}_{2,1\mathrm{V}}$	$\boldsymbol{X}_{2,1\mathrm{Q}}$	0.1270	0.0008	0.708	0	0.194	0.6601	1.876
$\boldsymbol{X}_{2,1\mathrm{V}}$	$\boldsymbol{X}_{2,1\mathrm{HR.150}}$	0.1276	0.0014	0.710	0.002	0.327	0.5681	1.870
$\boldsymbol{X}_{2,1\mathrm{V}}$	$\boldsymbol{X}_{2,1\mathrm{L}}$	0.1270	0.0008	0.708	0	0.187	0.6659	1.876
$\boldsymbol{X}_{2,2\mathrm{HF.612}}$	none	*0.0989*	–	*0.655*	–	–	–	*1.364*
$\boldsymbol{X}_{2,2\mathrm{HF.612}}$	$\boldsymbol{X}_{2,2\mathrm{Q}}$	0.0989	0.0001	0.660	0.005	0.016	0.8995	1.394
$\boldsymbol{X}_{2,2\mathrm{HF.612}}$	$\boldsymbol{X}_{2,2\mathrm{V}}$	0.0989	0.0001	0.662	0.007	0.014	0.9059	1.411
$\boldsymbol{X}_{2,2\mathrm{HF.612}}$	$\boldsymbol{X}_{2,2\mathrm{T.700}}$	0.0993	0.0005	0.655	0	0.108	0.7428	1.365
$\boldsymbol{X}_{2,2\mathrm{HF.612}}$	$\boldsymbol{X}_{2,2\mathrm{L}}$	0.0990	0.0002	0.660	0.005	0.043	0.8359	1.423
$\boldsymbol{X}_{2,3\mathrm{L}}$	none	*0.0312*	–	*0.602*	–	–	–	*1.911*
$\boldsymbol{X}_{2,3\mathrm{L}}$	$\boldsymbol{X}_{2,3\mathrm{HR.445}}$	0.0328	0.0016	0.602	0	0.343	0.5588	2.185
$\boldsymbol{X}_{2,3\mathrm{L}}$	$\boldsymbol{X}_{2,3\mathrm{Q}}$	0.0316	0.0004	0.602	0	0.083	0.7736	2.037



	Man2	Man	Auto|L	Auto|All	Auto|L|Xval
Man	0.9293				
Auto|L	0.4338	0.5357			
Auto|All	0.8883	0.9580	0.5614		
Auto|L|Xval	0.4372	0.5379	0.9999	0.5624	
Auto|All|Xval	0.8889	0.9550	0.5638	0.9986	0.5649

Fig. 15. Matrix of Kendall's rank correlation coefficients τ between vectors of predictions from the six Final MaxEnt models for example data set 2 (n = 256). Model characteristics are given in Table 14. All τ correspond to p values < 10^{-10}.

Table 14. Example data set 2: selection of MaxEnt models by the manual procedure for forward stepwise selection of parsimonious sets of derived variables (DVs) for each explanatory variable (EV), as outlined in Table 4, using the *F*-ratio test with significance level α = 0.05 as model improvement criterion. DVs are coded by type and identity of the EV from which they were derived in accordance with Table 2. V_t = fraction of total variation accounted for by the model; ΔV_t = fraction of total variation accounted for by the added DV; AUC_{corr} = AUC value, corrected for use with PO data; $F_{\text{df1,df2}}$, *pF* = *F* statistic and associated *p*-value, respectively, for *F*-ratio tests for comparison of models with a reference model; df1 and df2 denote the number of degrees of freedom for the numerator and the denominator, respectively, which are 1 and 204 with one-variable reference models and 1 and 203 with two-variable reference models; $\dot{q}_{\max}$ = maximum probability-ratio output value predicted by the model. Properties of the best one-variable reference model are shown on gray background.

DV in reference model	DV added	V_t	ΔV_t	AUC_{corr}	$F_{\text{df1,df2}}$	*pF*	$\dot{q}_{\max}$
$\boldsymbol{X}_{2,1\text{V}}$	none	0.1262	–	0.708	–	–	1.889
$\boldsymbol{X}_{2,1\text{V}}$	$\boldsymbol{X}_{2,2\text{HF}.612}$	0.2250	0.0989	0.788	26.022	<0.0001	2.564
$\boldsymbol{X}_{2,1\text{V}}$	$\boldsymbol{X}_{2,3\text{L}}$	0.1555	0.0293	0.747	7.082	0.0084	3.125
$\boldsymbol{X}_{2,1\text{V}}+\boldsymbol{X}_{2,2\text{HF}.612}$	$\boldsymbol{X}_{2,3\text{L}}$	**0.2435**	**0.0185**	**0.808**	**3.761**	0.0249	3.700
$\boldsymbol{X}_{2,1\text{V}}+\boldsymbol{X}_{2,2\text{HF}.612}$	$\boldsymbol{X}_{2,12\text{P}}$	0.2256	0.0006	0.788	0.125	0.7240	2.691
$\boldsymbol{X}_{2,1\text{V}}+\boldsymbol{X}_{2,2\text{HF}.612}$	$\boldsymbol{X}_{2,12\text{O}}$	0.2278	0.0027	0.790	0.558	0.4559	2.600

to the Man model except for not including variable $\boldsymbol{X}_{2,3\text{L}}$, is included in the comparison. The final (Man) model obtained by manual forward selection contained three DVs while four DVs were included in the Auto|L model and 11 DVs were included in the Auto|All model (Table 16).

The Man and Auto|All models were similar with respect to fractions of total variation accounted for and AUC_{corr} values; V_{Man} = 0.2435 and $V_{\text{Auto|All}}$ = 0.2688 and $\text{AUC}_{\text{corr,Man}}$ = 0.808 and $\text{AUC}_{\text{corr,Auto|All}}$ = 0.814, respectively (Table 16). Considerably lower fractions of total variation accounted for as well as AUC were observed for models with L variables as input than for models with all variables as input. Vectors of model predictions were more similar between the Man and Auto|All models (Kendall's τ = 0.958, *n* = 256) than between the Man and Man2 models (Kendall's τ = 0.929, *n* = 256; Fig. 15).

Predictions from the Auto|L models, which did not include the $\boldsymbol{X}_{2,1\text{V}}$ variable or any another variable that might open for modelling of a unimodal ecological response, differed strongly from predictions from all other models (Kendall's τ < 0.6 in all pairwise comparisons, *n* = 256; Fig. 15). This was reflected in different shapes of modelled ecological response curves for Sp2 with respect to the EV $\boldsymbol{Z}_{2,1}$ (Figs 16, 17a): by the Auto|L model, in which $\boldsymbol{Z}_{2,1}$ was represented by the L variable $\boldsymbol{X}_{2,1\text{L}}$, a linear response was modelled (Fig. 16c, 17a), while by the three models that included $\boldsymbol{X}_{2,1\text{V}}$, a truncated unimodel ('plateau-shaped') response was modelled (Figs 16a,b,d, 17a). Ecological response curves obtained by use of the Man2 model, with respect to both EVs $\boldsymbol{Z}_{2,1}$ and $\boldsymbol{Z}_{2,2}$, differed from response curves obtained by all other models by being smooth (Fig. 16a, 17). Ecological response curves obtained by all other models, which included the $\boldsymbol{X}_{2,3\text{L}}$ variable, had characteristic, 'saw-toothed' appearance due to systematic co-variation between

Table 15. Example data set 2: properties of final Maxent models, parameterised by the 'standard Maxent procedure', i.e., automated selection of derived variables (DVs) and ℓ_1-regularisation with default settings. Model char. = Model characteristics: Auto = model parameterised by the 'standard Maxent procedure' with default options and settings, including ℓ_1-regularisation; L = model parameterised by use of all L-type DVs, derived from the five explanatory variables (EVs); All = model parameterised by use of all DVs derived from all EVs by transformation outlined in Table 2; Xval = final model obtained by averaging five models obtained by 5-fold crossvalidation; DV # = number of DVs included in model, the identity of these DVs, coded by type and identity of the EV from which they were derived, is given in a footnote, for models obtained by crossvalidation the number of DVs in single models is given in brackets; $\acute{V}_t$ and V_t = regularised and unregularised fraction of total variation accounted for in model; AUC$_{\text{corr}}$ = AUC value, corrected for use with PO data; $F_{\text{df1,df2}}$ and pF are the value of the F statistic and the associated pF value, respectively, for an F-ratio test by which a MaxEnt model is compared with the null model, df1 and df2 denote the number of degrees of freedom for the numerator and the denominator, respectively; $\dot{q}_{\max}$ is the maximum predicted value, as given by the probability-ratio output format.

Model char.	DV #	V_t	V_t	AUC$_{\text{corr}}$	df1, df2	$F_{\text{df1,df2}}$	pF	$\dot{q}_{\max}$
Auto|L	4[1]	0.0964	0.1087	0.720	4, 201	6.128	0.0001	3.573
Auto|All	11[2]	0.2460	0.2688	0.814	11, 194	6.483	<0.0001	4.960
Auto|L|Xval	5[3] (5)	0.0856	0.1096	0.725	–	–	–	3.474
Auto|All|Xval	21[4] (9–15)	0.2320	0.2950	0.831	–	–	–	4.995

DVs: [1] $\boldsymbol{X}_{2,1\text{L}} + \boldsymbol{X}_{2,2\text{L}} + \boldsymbol{X}_{2,3\text{L}} + \boldsymbol{X}_{2,5\text{L}}$; [2] $\boldsymbol{X}_{2,1\text{Q}} + \boldsymbol{X}_{2,1\text{V}} + \boldsymbol{X}_{2,1\text{HR.150}} + \boldsymbol{X}_{2,1\text{HF.600}} + \boldsymbol{X}_{2,1\text{T.650}} + \boldsymbol{X}_{2,2\text{HF.612}} + \boldsymbol{X}_{2,3\text{Q}} + \boldsymbol{X}_{2,3\text{HR.445}} + \boldsymbol{X}_{2,4\text{Q}} + \boldsymbol{X}_{2,4\text{V}} + \boldsymbol{X}_{2,5\text{L}}$; [3] $\boldsymbol{X}_{2,1\text{L}} + \boldsymbol{X}_{2,2\text{L}} + \boldsymbol{X}_{2,3\text{L}} + \boldsymbol{X}_{2,4\text{L}} + \boldsymbol{X}_{2,5\text{L}}$;
[4] $\boldsymbol{X}_{2,1\text{L}} + \boldsymbol{X}_{2,1\text{Q}} + \boldsymbol{X}_{2,1\text{V}} + \boldsymbol{X}_{2,1\text{HR.150}} + \boldsymbol{X}_{2,1\text{HF.600}} + \boldsymbol{X}_{2,1\text{T.650}} + \boldsymbol{X}_{2,2\text{L}} + \boldsymbol{X}_{2,2\text{Q}} + \boldsymbol{X}_{2,2\text{V}} + \boldsymbol{X}_{2,2\text{HF.612}} + \boldsymbol{X}_{2,2\text{T.700}} + \boldsymbol{X}_{2,3\text{L}} + \boldsymbol{X}_{2,3\text{Q}} + \boldsymbol{X}_{2,3\text{V}} + \boldsymbol{X}_{2,3\text{HR.445}} + \boldsymbol{X}_{2,4\text{L}} + \boldsymbol{X}_{2,4\text{Q}} + \boldsymbol{X}_{2,4\text{V}} + \boldsymbol{X}_{2,5\text{L}} + \boldsymbol{X}_{2,5\text{Q}} + \boldsymbol{X}_{2,5\text{V}}$

$\boldsymbol{Z}_{2,3}$ and both of $\boldsymbol{Z}_{2,1}$ and $\boldsymbol{Z}_{2,2}$. Variation among neighbouring grid cells increased with increasing number of variables in the respective models, from the two-variable model Man2 (Fig. 16a) via the three-variable model Man (Fig. 16b) to the two Auto models (Figs 16c–d). Both Auto models included at least one of the random variables $\boldsymbol{Z}_{2,4}$ and $\boldsymbol{Z}_{2,5}$.

None of the six compared models gave rise to predictions that exceeded the probability ratio for presence sites in the saturated MaxEnt model of $\dot{q}_i = \frac{N}{n} = 5.333$, but values of $\dot{q}_i > 5.333$ were obtained for some grid cells in some of the single models that contributed to the cross-validated model (results not shown).

Results obtained by the four variable contribution measures were in good accordance (Table 16), expect for permutation importance (VC$_{\text{PI}}$) which in some cases deviated considerably from the other measures. This is exemplified by EVs $\boldsymbol{Z}_{2,1}$ and $\boldsymbol{Z}_{2,2}$ for the Auto|All model and for $\boldsymbol{Z}_{2,2}$ and $\boldsymbol{Z}_{2,3}$ for the Auto|L model. The tendency in example data set 1 for the AUC-based measure (VC$_{\text{AUC}}$) to put higher emphasis on contributions from individually less strongly significant variables was also observed for example data set 2. This accorded with the *relatively* much higher ΔAUC$_{\text{corr}}$ compared with the null model, than observed for V_t values for DVs derived from EVs $\boldsymbol{Z}_{2,3}$, $\boldsymbol{Z}_{2,4}$ and $\boldsymbol{Z}_{2,5}$ than from $\boldsymbol{Z}_{2,1}$ and $\boldsymbol{Z}_{2,2}$ (Table 12).

Results of manual forward stepwise model selection (Tables 12–14) and variation ac-

Table 16. Example data set 2: comparison of final MaxEnt models. Model char. = Model characteristics; Man2 = two-variable model parameterised by the manual procedure for forward stepwise selection of derived variables (DVs) and explanatory variables (EVs) outlined in Table 4, using the *F*-ratio test with significance level α = 0.05 as model improvement criterion; Man = final model with three DVs parameterised by manual forward stepwise selection; Auto = model parameterised by the 'standard Maxent procedure' with default options and settings, including ℓ_1-regularisation; L = model parameterised by use of L-type DVs derived from the five EVs; All = model parameterised by use of all DVs derived from all EVs by the procedure outlined in Table 2; Xval = final model obtained by averaging five models obtained by 5-fold crossvalidation; DV # = number of derived variables included in model, the identity of these derived variables, coded by type and identity of the EV from which they were derived, is given in a footnote, for models obtained by crossvalidation the number of EVs in single models is given in brackets; V_t, = (unregularised) fraction of total variation accounted for by model; AUC$_{\text{corr}}$ = AUC value, corrected for use with PO data; VC$_{\text{PI}}$, VC$_{\text{PC}}$, VC$_{\text{AUC}}$ and VC$_{\text{FVA}}$ = variable contributions calculated for each EV by four different measures (see text for explanation), expressed as fractions of the sum of contributions by all variables.

Model char.	DV #	V_t	AUC$_{\text{corr}}$	EV	VC$_{\text{PI}}$	VC$_{\text{PC}}$	VC$_{\text{AUC}}$	VC$_{\text{FVA}}$
Man2	2[1]	0.2250	0.788	$\boldsymbol{Z}_{2,1}$	0.506	0.560	0.573	0.561
				$\boldsymbol{Z}_{2,2}$	0.494	0.440	0.427	0.439
				$\boldsymbol{Z}_{2,3}$	–	–	–	–
				$\boldsymbol{Z}_{2,4}$	–	–	–	–
				$\boldsymbol{Z}_{2,5}$	–	–	–	–
Man	3[2]	0.2435	0.808	$\boldsymbol{Z}_{2,1}$	0.509	0.520	0.447	0.492
				$\boldsymbol{Z}_{2,2}$	0.475	0.402	0.333	0.386
				$\boldsymbol{Z}_{2,3}$	0.016	0.078	0.219	0.122
				$\boldsymbol{Z}_{2,4}$	–	–	–	–
				$\boldsymbol{Z}_{2,5}$	–	–	–	–
Auto|L	4[3]	0.1087	0.720	$\boldsymbol{Z}_{2,1}$	0.101	0.172	0.226	0.175
				$\boldsymbol{Z}_{2,2}$	0.744	0.570	0.377	0.497
				$\boldsymbol{Z}_{2,3}$	0.056	0.185	0.241	0.252
				$\boldsymbol{Z}_{2,4}$	–	–	–	–
				$\boldsymbol{Z}_{2,5}$	0.099	0.073	0.157	0.073
Auto|All	11[4]	0.2688	0.814	$\boldsymbol{Z}_{2,1}$	0.372	0.528	0.608	0.687
				$\boldsymbol{Z}_{2,2}$	0.548	0.383	0.145	0.189
				$\boldsymbol{Z}_{2,3}$	0.054	0.072	0.164	0.104
				$\boldsymbol{Z}_{2,4}$	0.017	0.001	0.021	0.002
				$\boldsymbol{Z}_{2,5}$	0.009	0.015	0.062	0.017
Auto|L|Xval	5[5] (5)	0.1096	0.725	$\boldsymbol{Z}_{2,1}$	0.144	0.162	0.220	0.174
				$\boldsymbol{Z}_{2,2}$	0.490	0.557	0.367	0.495
				$\boldsymbol{Z}_{2,3}$	0.283	0.191	0.234	0.252
				$\boldsymbol{Z}_{2,4}$	0.001	0.001	0.025	0.003
				$\boldsymbol{Z}_{2,5}$	0.083	0.089	0.157	0.073
Auto|All|Xval	21[6] (9–15)	0.2950	0.831	$\boldsymbol{Z}_{2,1}$	0.485	0.510	0.421	0.462
				$\boldsymbol{Z}_{2,2}$	0.427	0.368	0.389	0.434
				$\boldsymbol{Z}_{2,3}$	0.036	0.080	0.136	0.090
				$\boldsymbol{Z}_{2,4}$	0.017	0.016	0.017	0.002
				$\boldsymbol{Z}_{2,5}$	0.035	0.025	0.037	0.013

DVs: [1]$\boldsymbol{X}_{2,1\text{V}} + \boldsymbol{X}_{2,2\text{HF.612}}$; [2]$\boldsymbol{X}_{2,1\text{V}} + \boldsymbol{X}_{2,2\text{HF.612}} + \boldsymbol{X}_{2,3\text{L}}$; [3]$\boldsymbol{X}_{2,1\text{L}} + \boldsymbol{X}_{2,2\text{L}} + \boldsymbol{X}_{2,3\text{L}} + \boldsymbol{X}_{2,5\text{L}}$; [4]$\boldsymbol{X}_{2,1\text{Q}} + \boldsymbol{X}_{2,1\text{V}} + \boldsymbol{X}_{2,1\text{HR.150}} + \boldsymbol{X}_{2,1\text{HF.600}} + \boldsymbol{X}_{2,1\text{T.650}} + \boldsymbol{X}_{2,2\text{HF.612}} + \boldsymbol{X}_{2,3\text{Q}} + \boldsymbol{X}_{2,3\text{HR.445}} + \boldsymbol{X}_{2,4\text{Q}} + \boldsymbol{X}_{2,4\text{V}} + \boldsymbol{X}_{2,5\text{L}}$; [5]$\boldsymbol{X}_{2,1\text{L}} + \boldsymbol{X}_{2,2\text{L}} + \boldsymbol{X}_{2,3\text{L}} + \boldsymbol{X}_{2,4\text{L}} + \boldsymbol{X}_{2,5\text{L}}$; [6]$\boldsymbol{X}_{2,1\text{L}} + \boldsymbol{X}_{2,1\text{Q}} + \boldsymbol{X}_{2,1\text{V}} + \boldsymbol{X}_{2,1\text{HR.150}} + \boldsymbol{X}_{2,1\text{HF.600}} + \boldsymbol{X}_{2,1\text{T.650}} + \boldsymbol{X}_{2,2\text{L}} + \boldsymbol{X}_{2,2\text{Q}} + \boldsymbol{X}_{2,2\text{V}} + \boldsymbol{X}_{2,2\text{HF.612}} + \boldsymbol{X}_{2,2\text{T.700}} + \boldsymbol{X}_{2,3\text{L}} + \boldsymbol{X}_{2,3\text{Q}} + \boldsymbol{X}_{2,3\text{V}} + \boldsymbol{X}_{2,3\text{HR.445}} + \boldsymbol{X}_{2,4\text{L}} + \boldsymbol{X}_{2,4\text{Q}} + \boldsymbol{X}_{2,4\text{V}} + \boldsymbol{X}_{2,5\text{L}} + \boldsymbol{X}_{2,5\text{Q}} + \boldsymbol{X}_{2.}$

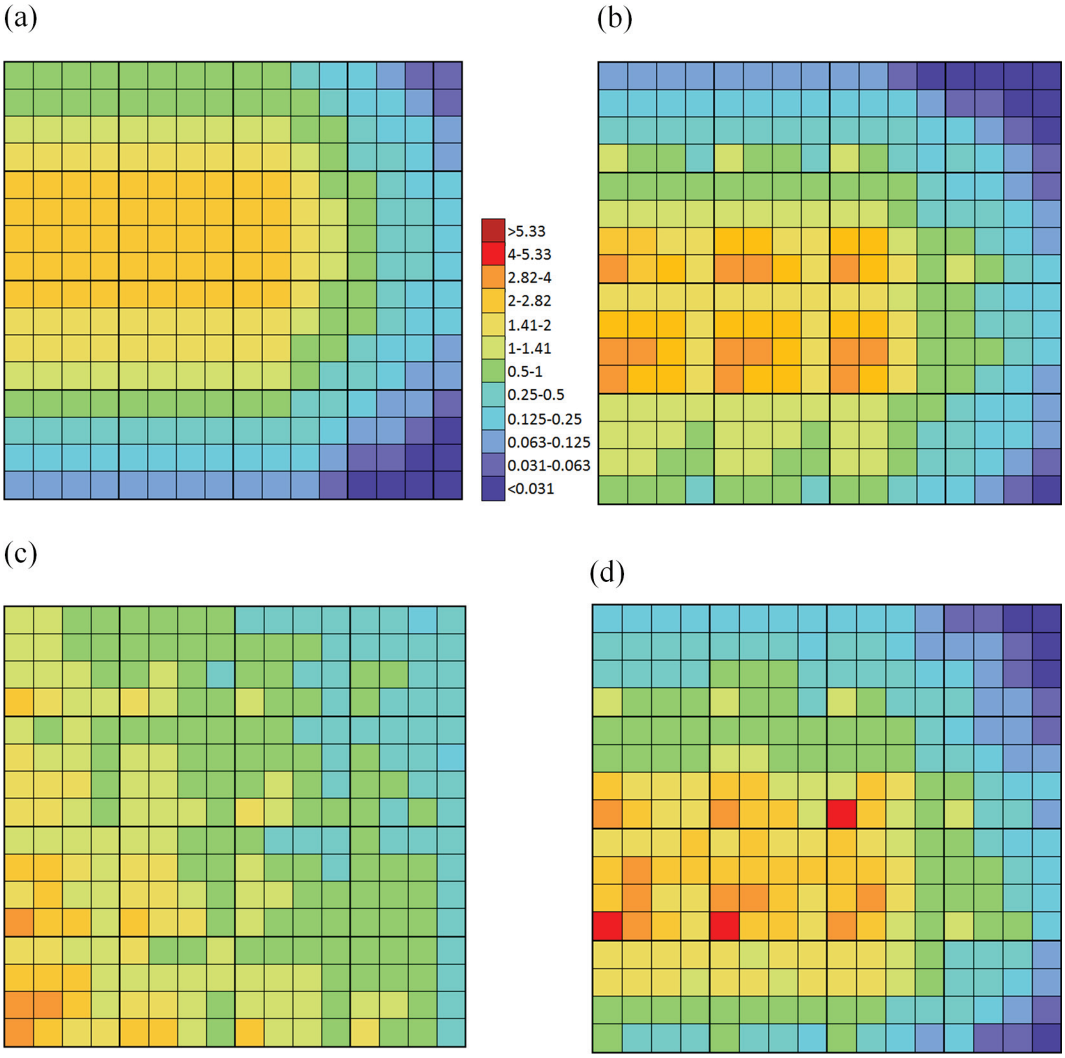


Fig. 16. Map representation of predictions for the modelled target Sp2 in example data set 2, given in probability-ratio output format $\dot{q}$. (a) The Man2 model, parameterized by the manual procedure for forward stepwise selection of derived variables (DVs) and explanatory variables (EVs) outlined in Table 4, using the *F*-ratio test with significance level $\alpha = 0.05$ as model improvement criterion. Two DVs are included in the model. (b) The Man model, the final model with three DVs, obtained by the manual forward selction procedure. (c) The Auto|L model, parameterized by the 'standard Maxent procedure' with default options and settings, including ℓ_1-regularisation, by use of all L-type DVs derived from the five EVs. (d) The Auto|All model, parameterized by the 'standard Maxent procedure', by use of all DVs derived from all EVs by transformations outlined in Table 2.

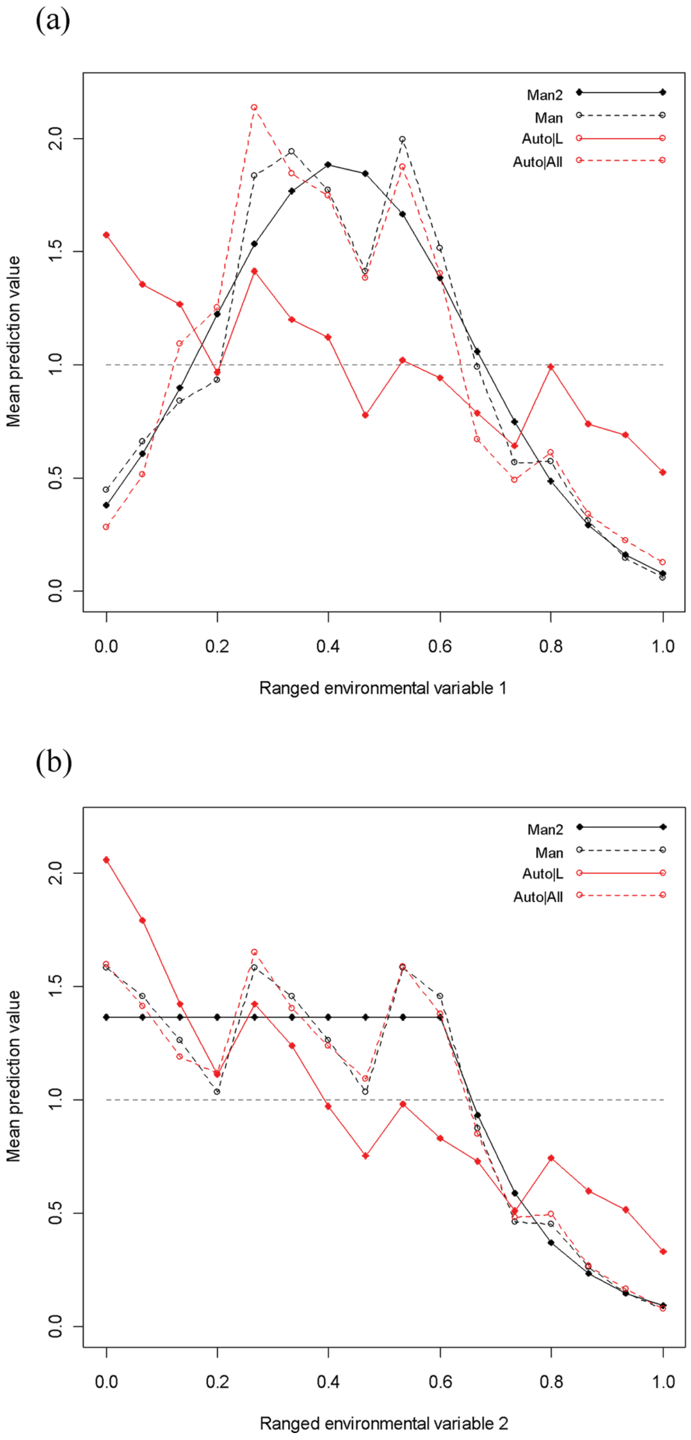


Fig. 17. Overall ecological response curves for the modelled target SP2 in example data set 2, given in probability-ratio output format $\dot{q}$. (a) Response to environmental variable (EV) $\boldsymbol{Z}_{2,1}$. Which equals DV $\boldsymbol{X}_{2,1L}$. (b) Response to environmental variable (EV) $\boldsymbol{Z}_{2,2}$, which equals DV $\boldsymbol{X}_{2,2L}$.

counted for (v_t) by the three DVs in the Man model (blue bars in Fig. 18a; which correspond to V_t values in Table 12 multiplied with V_S = 1.674) accorded well with results for 'jackknife variation accounted for', obtained by leaving out the variable in question from the model (tan-coloured bars in Fig. 18a): all three DVs added to the explanatory power of the model, in order of decreasing importance $\boldsymbol{X}_{2,1\mathrm{V}} > \boldsymbol{X}_{2,2\mathrm{HF.612}} > \boldsymbol{X}_{2,3\mathrm{L}}$. The corresponding results for the Auto|All model (Fig. 18b) were less clearly interpreted although inclusion of $\boldsymbol{X}_{2,1\mathrm{HR.150}}$ and $\boldsymbol{X}_{2,2\mathrm{HF.612}}$ in the final model was justified by these being the only DVs that accounted for variation that was not also accounted for by other variables (slightly shorter tan bars than other derived variables in Fig. 18b) and by their larger regularised variation accounted for $\acute{v}_t$ (longer blue bars; Fig. 18b). Of the five DVs with 0.10 < $\acute{v}_t$ < 0.15, only the one ($\boldsymbol{X}_{2,1\mathrm{T.650}}$) which accounted for the largest fraction of total variation in addition to the best DV derived from the same EV (see Table 13) was included in the Auto|All model. Neither single-variable variation accounted for (Table 12) nor independent contribution to the model (Table 13) could, however, explain why the DV $\boldsymbol{X}_{2,1\mathrm{HR.150}}$ was included in the model, why EV $\boldsymbol{Z}_{2,1}$ was represented by four DVs in the final model, why $\boldsymbol{X}_{2,3\mathrm{HR.445}}$ was preferred over $\boldsymbol{X}_{2,3\mathrm{L}}$, why EVs $\boldsymbol{Z}_{2,4}$ and $\boldsymbol{Z}_{2,5}$ were included, nor why $\boldsymbol{Z}_{2,4}$ was represented by two variables, $\boldsymbol{X}_{2,4\mathrm{Q}}$ and $\boldsymbol{X}_{2,4\mathrm{V}}$.

DISCUSSION

CHOICE AMONG TYPES OF DERIVED VARIABLES

The results obtained for fractions of total variation accounted for by, and frequency-of-observed-presence curves with respect to, DVs of different types derived from the same EV, exemplify two typical properties of DVs that contribute strongly to MaxEnt models: (1) ability to concentrate presence grid cells to a narrow interval near one end of the [0, 1] range of DV values or, equivalently, to make the mean value for the DV in observed presence cells maximally different from the mean in uninformed background cells; and (2) high maximum frequency of observed presence. From the perspective of (1), the ideal variable, which explains all variation in the response of the modelled target, is a threshold-type DV that separates the grid cells into one group of observed presence cells and one group of uninformed background cells. Threshold-shaped ecological responses, i.e., large but predictable response to small changes in an explanatory variable, are, however, likely to be very rare (Halvorsen 2012). The gradient analytic perspective predicts smooth overall ecological response curves to important environmental complex-gradients, that level off gradually from a mode (optimum) towards the modelled target's tolerance limits [see Halvorsen (2012) and references cited therein]. The fact that this principle was used to construct example data sets in this study explains the absence of cases where a threshold-type DV explains more variation than the best DV of other types. The importance of property (1) is also shown by DVs derived from EVs $\boldsymbol{Z}_{1,1}$ in example data set 1 and $\boldsymbol{Z}_{2,1}$ and $\boldsymbol{Z}_{2,2}$ in example data set 2. Among DVs derived from $\boldsymbol{Z}_{1,1}$, $\boldsymbol{X}_{1,1\mathrm{L}}$ is least successful in concentrating observed presence observations to a narrow interval, and hence explains less variation than $\boldsymbol{X}_{1,1\mathrm{V}}$ and $\boldsymbol{X}_{1,1\mathrm{Q}}$ (compare Fig. 9 and Table 8). Furthermore, among DVs derived from EV $\boldsymbol{Z}_{2,1}$, V_t increases sixfold from the L via the Q to the V variable, corresponding to stronger concentration of presence grid cells to low variable values (compare Fig. 13a and Table 12). For $\boldsymbol{Z}_{2,1}$, the ability to concentrate observed presence grid cells is accompanied by a displacement of the peak of the frequency-of-observed-presence

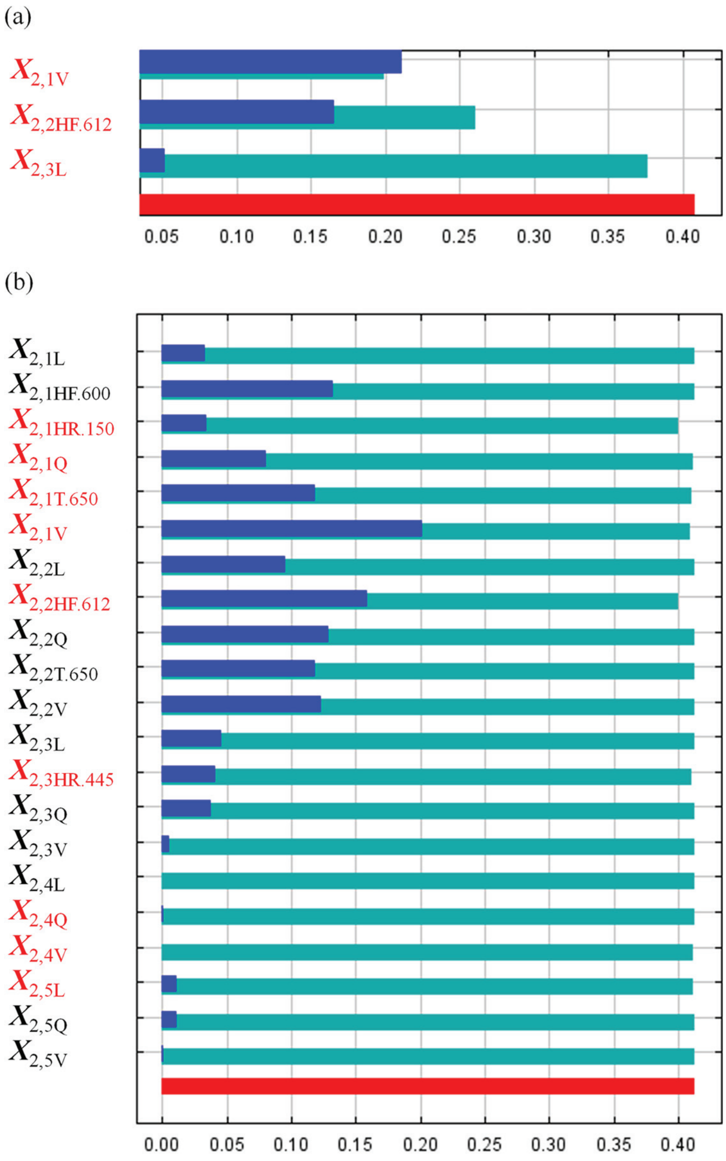


Fig. 18. Example data set 2: contributions from single derived variables (DVs) to MaxEnt models, given as 'jackknife' estimates from Maxent software. The performance criterion, expressed on the horizontal axis, is variation accounted for by the model with the DV in question left out, as shown by tan-coloured bars, and variation accounted for by the single-variable model for the DV in question. The variation accounted for by the full model is shown by the red bar. (a) The Man model, parameterized by the manual procedure for forward stepwise selection of derived variables (DVs) and explanatory variables (EVs) outlined in Table 4, using the *F*-ratio test with significance level α = 0.05 as model improvement criterion. Values on the horizontal axis are unregularised variation accounted for, V_t. (b) The Auto|All model, parameterized by the 'standard Maxent procedure' with default options and settings, including ℓ_1-regularisation, by use of all DVs derived from all EVs by transformations outlined in Table 2. Values on horizontal axis are regularised variation accounted for, $\acute{V}_t$.

curves towards a value of 0 for the DV, which results in an ecological response that is closer to monotonous. For DVs derived from $\boldsymbol{Z}_{2,2}$, frequency-of-observed-presence patterns as well as the increase in V_t from the L via the Q to the HF variable (Fig. 13b) show that concentration of observed presence grid cells to a narrow interval along the ranged variable is accompanied by a more monotonous response in the range spanned by observed presence grid cells. The importance of property (2) is demonstrated by the lower V_t of $\boldsymbol{X}_{1,1\mathrm{V}}$ than of $\boldsymbol{X}_{1,1\mathrm{Q}}$, and by the lower V_t of $\boldsymbol{X}_{2,1\mathrm{HF}.600}$ than of $\boldsymbol{X}_{2,1\mathrm{V}}$.

Single-variable MaxEnt models show that choice of DV type of derived variable strongly influences MaxEnt model performance: variation in the fraction of total variation accounted for by a factor of up to 6 (for $\boldsymbol{Z}_{2,1}$) is observed between DVs of different types derived from the same EV. No DV-type performs generally best, but T-type variables are never among the best-performing DVs: variables of the continuous or other spline types have the best predictive ability in at least one case; the L variable $\boldsymbol{X}_{2,3\mathrm{L}}$ for $\boldsymbol{Z}_{2,3}$, the Q variable $\boldsymbol{X}_{1,1\mathrm{Q}}$ for $\boldsymbol{Z}_{1,1}$, the V variable $\boldsymbol{X}_{2,1\mathrm{V}}$ for $\boldsymbol{Z}_{2,1,}$ and the hinge-type variable $\boldsymbol{X}_{2,2\mathrm{HF}.612}$ for $\boldsymbol{Z}_{2,2}$. The DV that best separates observed presence grid cells from uninformed background cells explains most variation in MaxEnt models. For modelled targets that respond monotonously to environmental complex-gradients the shape of the frequency-of-observed-presence curve determines which type of monotonous transformation that gives the best result. Conversely, the best-performing type of DV in each case is to some extent predictable from frequency-of-observed-presence curves. This result suggests that restricting oneselves to one type of DV, e.g., hinge variables, as done by, e.g., Thompson et al. (2011), is not recommended.

The far better performance of the variance variable $\boldsymbol{X}_{2,1\mathrm{V}}$ than of variables of any other type derived from $\boldsymbol{Z}_{2,1}$ clearly shows that appropriate modelling of targets with unimodal response to important environmental gradients requires DV of the deviation (D) type, i.e., DVs that express ‘distance from optimum’. Even if the $\boldsymbol{X}_{2,1\mathrm{V}}$ variable performed best among the six transformations of $\boldsymbol{Z}_{2,1}$ in example data set 2 (Table 12), this DV was unable to concentrate observed presence grid cells to values of the DV close to 0. The reason for this is that the simple transformation into the V variable does not take into account that the modelled target’s overall ecological response curve is truncated on one side (Fig. 6b). More variation is therefore likely to be accounted for by a DV constructed by first estimating the target’s optimum and then using this optimum instead of the mean value of the derived variable in observed presence cells to construct a deviation-type variable.

The strong improvement of models resulting from transformation of L variables into the Q variable, which is observed for $\boldsymbol{Z}_{1,1}$, $\boldsymbol{Z}_{2,1}$ and $\boldsymbol{Z}_{2,3}$, suggests that cases are likely to exist in which other monotonous transformations of EV than the arbitrarily chosen Q variable, will improve MaxEnt models considerably. This accords with the fundamental insight from gradient analysis that species do not necessarily respond linearly to environmental gradients scaled in physical or chemical units (e.g., Økland 1990, 1992), and that the scaling of environmental gradients therefore, essentially, is arbitrary (Minchin 1989). This has the important implication for MaxEnt modelling that, ideally, the modeller should search for the realistic, i.e., simple, monotonous transformation of each explanatory variable that maximises the fraction of the total variation accounted for. This can be done by procedures like the V_t-knot graph approach adopted in this paper for tuning of DV of the hinge and threshold types. One family of monotonous transformation functions that may suit this purpose is the zero-skewness transformation (Økland et al. 2001, 2003), which is used to manipulate skewness in EVs prior to statistical analyses by GLM (e.g., Bakkestuen et al. 2009, Rydgren et al. in press). The zero-skewness transformation implies that right-skewed variables are transformed by the function ln $(c + z)$ and left-skewed variables by e^{cz}; in both cases the value of the scalar c is determined so that the skewness is zero.

The failure of the simple V transformation of the $\boldsymbol{Z}_{2,1}$ variable to provide a monotonous 'distance from optimum' function suggests that more complex D-type transformations are needed to account for a realistic range of ecological responses of modelled targets. This accords with recommendations in several distribution modelling studies by logistic regression, to consider carefully which transformation is likely to optimise the model's predictive ability (e.g., Santika & Hutchinson 2009, Gastón & García-Viñas 2011, Michel et al. 2011). Empirical evidence on ecological response curves (Oksanen & Minchin 2002, Rydgren et al. 2003), e.g., obtained by the HOF (Huisman-Olff-Fresco) modelling framework (Huisman et al. 1993, Oksanen & Minchin 2002), indicates that a function with four parameters (four degrees of freedom) is generally sufficient to capture generalisable patterns of variation in species' aggregated performance (Halvorsen 2012) along gradients. Further research is needed to optimise construction of parsimonious sets of derived variables for MaxEnt modelling in ways that account for unimodality, skewness and/or platy- or leptokurtosis in frequency-of-observed-presence curves for the modelled targets.

MODEL SELECTION

The automated, standardised procedure for formation of DVs and model selection implemented as default in Maxent software, here referred to as 'standard Maxent practice', fails to return adequate models for example data set 2, both when the five EVs are represented by L variables (the Auto|L model) and when the full set of 21 DVs manually derived from the five EVs are used as input to the Maxent software (the Auto|All model). Poor performance of the Auto|L model is clearly demonstrated by the low $\mathrm{AUC}_{\mathrm{corr}}$ value and the low fraction of variation accounted for by this model, compared with other models (Tables 15–16). The Auto|L model fails on two points: (1) it is mis-specified; and (2) it is overfitted to the data used to parameterise the model. Signs of mis-specification are: the failure to predict the unimodal response to $\boldsymbol{Z}_{2,1}$, as clearly shown by the ecological response curve (Fig. 17a) and the map representation of predictions (Fig. 16c), and the low variable contributions attributed to $\boldsymbol{Z}_{2,1}$ (Table 16) relative to other models. Mis-specification is a consequence of only four L variables being included in the model. Two pathways for modelling unimodal responses are, in principle, available in Maxent software when $n = 48$ (Phillips et al. 2006): (1) to combine L and Q variables, a Q variable is taken into consideration when the number of observed presences, $n \geq 10$ (Phillips et al. 2006); and (2) to combine two or more hinge (and/or threshold) variables, hinge variables are taken into consideration when $n \geq 15$ (Phillips et al. 2006). None of these pathways are activated in this case. Even though the mis-specification problem of the Auto|L model is a result of the model's simplicity, the model is at the same time overfitted to the data. Indications that the Auto|L model is overfitted to the data are: (1) That the random explanatory variable $\boldsymbol{X}_{2,5\mathrm{L}}$ is included in the model, and attributed a contribution of 7.3–15.7 %, depending on the measure used to quantify variable contribution. Model predictions (Fig. 16c) therefore reflect variation in random derived variables to a degree that is visible on relevant graphs (cf. Figs 6a, 6c). (2) That the model, despite having more parameters than the Man model (4 vs 3), accounts for less than half of the variation accounted for by the latter. Accordingly, the Auto|L model is likely to suffer from Type I overfitting, , as defined by Halvorsen (2012), that a more complex model has lower predictive performance (on independent data) than a simpler model.

Comparisons between the Auto|All and Man models indicate that also the former may suffer from overfitting of Type II, i.e., that a more complex model is similar in predictive performance than a more complex model. Indications of Type II overfitting in the Auto|All model are: (1) considerable local variability of predictions (see prediction map in Fig. 16d); (2) similar $\mathrm{AUC}_{\mathrm{corr}}$ values and fractions of total variation accounted for by the Auto|All model with 11 DVs and the Man model with only three DVs; and (3) that eight of the 11 DVs in the Auto|All model do not make individually significant contributions to explaining variation in observed presence of the modelled target in single-variable tests. However, since the simulated example data sets 1 and 2 used in this paper are of the PO type and not sampled from an underlying, known distribution of presence and absence observations, the ultimate test for overfitting, performance on independent data, is not applicable. Type II overfitting can therefore not be conclusively demonstrated in this case.

These results show that the best MaxEnt model is not simply the model with optimal complexity, but instead indicate that model complexity is itself a complex matter that cannot be represented along one linear gradient. One important aspect of model complexity is *model specification*, i.e., the extent to which the modelled response to an important environmental gradient has an appropriate curve shape. The discussion in the previous section shows that response-curve shape is controlled by the process by which parsimonious sets of derived variables are constructed from each explanatory variable. The other, in itself very complex, issue, is to find the optimal level of model complexity *sensu stricto*, i.e., the appropriate number of DVs to be included in the model and best possible parameter estimates.

No universally 'right' complexity level exists, not even for a given modelled target in a given study area: the purpose of the DM study is a main determinant of which complexity level is the most appropriate (Barry & Elith 2006, Elith et al. 2010, Halvorsen 2012). Halvorsen (2012) argues that, from a gradient analytic perspective, modelling purposes can be divided into two main groups according to applicability of different methods and approaches for model performance assessment: (1) ecological response modelling (ERM) and projective distribution modelling (PPM); and (2) spatial prediction modelling (SPM). The most appropriate distribution models for the ERM and PPM purposes summarise relationships that are valid over most of, or the entire, distribution area of the modelled target. Good ERM (and PPM) models should therefore be simple in terms of number of EVs and DVs included. Because the true species-environment relationship is an ideal which can never be modelled correctly in all details, no truth, possible to represent by empirical data, normally exists against which ERM and PPM models can be evaluated. SPM models, on the other hand, should be evaluated pragmatically by comparing their predictive performance on truly independent evaluation data. I refer to Halvorsen (2012) and references quoted therein (e.g., Araujo & Guisan 2006, Lahoz-Monfort et al. 2007, Veloz 2009, Edrén et al. 2010, Warren & Seifert 2011) for discussions of the importance of using of independent P/A data for evaluation of SPM models. SPM calls for a level of complexity that matches the complexity of variation in the study area (Warren & Seifert 2011).

Because good spatial predictions are targeted in a large majority of empirical DM studies (Franklin 2009), the following discussion mainly addresses model selection in the SPM context. Predictive performance in the study area is also the main focus in most comparative studies of DM methods (e.g., Elith et al. 2006), targeted in experiments for tuning of the 'standard Maxent procedure' by Phillips & Dudík (2008), and addressed in most other studies in which model selection in MaxEnt is discussed (e.g., Anderson & Gonzalez 2011, Phillips 2011, Warren & Seifert 2011). The results obtained for example data set 2 have particular relevance for model selection in MaxEnt, used for the SPM purpose. Both of the Auto|L and Auto|All models for example data set 2 are overfitted because random variables are not de-selected or sufficiently strongly downweighted by the 'standard Maxent procedure' or, put in other words, because the ℓ_1-regularisation procedure with automated settings imposes a regularisation that is too weak. A relevant perspective on this issue is that the regulation parameters λ, given by expression (52), are < 0.15 for all 11 derived variables included in the Auto|All model (results not shown) while the *F*-ratio test with $\alpha = 0.05$ corresponds to $\lambda = 3.841$.

The obvious way to reduce the danger of overfitting is to apply a stronger regularisation. This has been suggested by several authors (e.g., Phillips & Dudík 2008, Elith et al. 2010, Anderson & Gonzalez 2011). Thus Lamb et al. (2008) use λ = 2.5 'to account for statistical overfitting given the relatively large number of predictors', Naimi et al. (2011) use λ = 2.5 to obtain 'predicted response shapes [that are] visually closest to the ones used to simulate data sets' and Warren & Seifert (2011) find that optimal regularisation values are generally higher than Maxent default values. Also Anderson & Gonzalez (2011), in their study of the rare shrew *Cryptotis meridensis*, find optimal regularisation values different from, and usually higher than, Maxent default values. Anderson & Gonzalez (2011) conclude that generally applicable default values for regularisation parameters are likely not to be found because of strong idiosyncrasies in the properties of modelled targets, e.g., species. Instead, they advocate species-specific tuning of regularisation parameters.

A closer look at the basis for the variable-type specific default regularisation values λ_K in the Maxent software (Phillips & Dudík 2008) reveals that these values have a weak empirical basis despite based upon numerous parallel runs for many different data sets. Phillips & Dudík (2008: Fig. 2, lower panel) use log loss and AUC calculated by data-splitting and five-fold crossvalidation for internal model performance assessment. It is not clear from their paper if unregularised or regularised log loss was used. Regularised log loss, if used, is not a measure of variation *as such*, but a model optimisation criterion of the penalised likelihood type, like AIC and BIC. Regularised log loss is therefore not comparable between models parameterised on different *n* or by use of different regularisation parameters. AUC, on the other hand, is comparable among models. Inspection of the 43 curves in the paper by Phillips & Dudík (2008: Fig. 2), obtained for random subsets with different *n*, subsampled from 12 different data sets, with AUC as performance statistic, shows 17 curves for which AUC decreases monotonously with λ_K, 10 that are nearly flat, and 16 that are unimodal or monotonously increasing. Monotonously decreasing curves accord with best performance without regularisation, i.e., for $\lambda_K = 0$, flat curves accord with model performance independent of the strength of regularisation, while monotonously increasing (or unimodal) curves indicate existence of a value $\lambda_K > 0$ at which model performance is better than in models without regularisation. These disparate results show that the effect of regularisation is strongly context-dependent and that reliance on pre-tuned regularisation parameters is likely to give rise to suboptimal models in many cases. Similar views on regularisation by shrinkage methods are expressed in a more general context by, e.g., Hastie et al. (2009).

Let us consider the typical situation that DM is performed for the SPM purpose, and that models cannot be evaluated by the ultimate performance measure, predictive ability on independent P/A data, because such data are not available. For such cases, *a priori* choices of model selection strategy, internal model performance measure(s), and model improvement criterion, have to be made. Two options are available: (1) to adopt a 'consensus MaxEnt practice' that is considered as base based upon extensive comparative studies; or (2) to apply a set of rules for tuning of MaxEnt settings and options based upon specific knowledge about properties of the modelled target, the study area, and/or the data sets to be used in the modelling. These rules will be referred to as 'best specific MaxEnt practice'. Finding such a best specific MaxEnt practice is an attractive goal from a theoretical point of view, but I agree with Anderson & Gonzalez (2011) that rules for data-driven tuning of MaxEnt options and settings will be hard or perhaps impossible to find. This pessimism reflects the fact that no characteristic of modelled targets, deducible from PO data, has so far turned up that can be linked directly to regularisation settings. If no best specific MaxEnt practice for tuning of regularisation parameters can be found, ℓ_1-regularisation will remain burdened with strong elements of unpredictability, due to idiosyncratic properties of the modelled target as well as of the PO data set, and subjectivity, due to the need for reliance on pre-set regularisation parameters.

Another major problem with shrinkage methods, which has largely been neglected in discussions of ℓ_1-regularisation in MaxEnt, is that stronger regularisation does not only raise the threshold for inclusion of derived variables in the model but at the same time increases model bias (Reineking & Schröder 2006). The effect of increasing bias is illustrated by the observation of Warren et al. (2011) that very strong regularisation results in models with no other parameters than the intercept: among realistic models, the null model is the maximally biased model. Strong regularisation is therefore at odds with the fundamental principle of statistical modelling, that estimators should be unbiased (e.g., Sokal & Rohlf 1995). This leaves us with the basic question: which model selection approach does, in general, result in distribution models with best predictive performance?

The worked examples indicate that shrinkage methods do not necessarily result in distribution models with better predictive performance than subset selection methods. To the contrary, the results indicate that careful manual forward stepward selection, first to produce a parsimonious set of DVs from each EV, and thereafter to build a final model from these parsimonious sets (see Table 4), may provide the control over model complexity needed to obtain models of adequate fit. This contradicts the current paradigm in MaxEnt modelling, that ℓ_1-regularisation is one of the major reasons why MaxEnt consistently performs among the best DM methods. While MaxEnt has been compared with other methods in many studies (see the introduction chapter ' MaxEnt modelling of distributions'), the relative performance of MaxEnt with different model selection methods remains incompletely explored. The tuning of MaxEnt regularisation parameters by Phillips & Dudík (2008) was performed with obviously overfitted models, with all regularisation parameters $\lambda = 0$, as a reference.

The results of simple worked examples in this study do not give strong reasons to claim that MaxEnt models with subset selection of derived variables will be consistently better than models that make use of ℓ_1-regularisation [or other shrinkage methods; see Dudík et al. (2007)]. Furthermore, they definitively do not prove that the manual forward stepwise selection procedure proposed in this paper (see Table 4) is optimal among subset selection methods. In fact, results obtained for example data set 1 indicate that better models may be obtained by a more flexible approach, such as forward-backward selection. The results do, however, show that choice of model selection method is so crucial for the performance of modelling methods, MaxEnt included, that the hypothesis that MaxEnt models often or in most cases perform better with settings other than those of the 'standard Maxent practice' clearly needs to be further explored. Studies with the aim of assessing the relative merits of model selection approaches should: (1) compare a realistic range of settings for all model selection methods; (2) use independent P/A evaluation data for assessment of the models' relative predictive performance; and (3) make use of study systems that differ with respect to modelled targets, ecosystems, geographical areas, grains and extents, and explanatory variables.

Results of MaxEnt modelling with different sets of derived variables, obtained for the two simple simulated data sets in this study, suggest that the ability of MaxEnt models to explain variation in the distribution of a modelled target may be enhanced, without inappropriate increase of model complexity, by manual control over the process by which explanatory variables are transformed into derived variables and the latter are selected for inclusion in the final model. Manual pre-selection of EVs has been reported to give favourable results in many DM studies, by MaxEnt (e.g., Santika & Hutchinson 2009, Wollan et al. 2008) as well as by other modelling methods (e.g., Pearce & Ferrier 2000a, Suárez-Seoane et al. 2004, Wohlgemuth et al. 2008, Platts et al. 2010). Furthermore, the results of this study suggest that selection of DVs should be guided by patterns of variation in frequency of observed presence of the modelled target with respect to the EVs in question.

This study shows that manual stepwise subset selection of DVs and EVs results in simpler models with fewer parameters than models obtained by shrinkage methods. Simpler models in terms of number of parameters have the additional advantage of being more easily interpretable than more complex models (Buermann et al. 2008, Parolo et al. 2008, Wollan et al. 2008, Warren & Seifert 2011) and may therefore be more useful for understanding which factors are responsible for the observed distributions (Austin 2007, Halvorsen 2012). This opens for the possibility that manually built MaxEnt models are more likely to express patterns that are so general that they may serve the ERM purpose and be useful for the PPM purpose while at the same time not compromising the demand of SPM models for high predictive performance.

METHODS AND APPROACHES FOR INTERNAL MODEL PERFORMANCE ASSESSMENT AND CHOICE OF MODEL IMPROVEMENT CRITERION

Choice of model improvement criterion is tightly coupled with choice of method or approach for internal model performance assessment. Two main types of model improvement criteria are currently in use: (1) a threshold value for a performance statistic; or (2) a significance level α for rejection of an appropriate null hypothesis which typically is that addition of a variable or several variables to a model does not improve the model significantly more than expected of a random variable. Typically, the statistical test by which (2) is accomplished takes the numbers of observed presence and uninformed background observations explicitly into account while this is not necessarily the case for threshold values for performance statistics used directly as model improvement criterion. Threshold values can be set subjectively, as exemplified by Wollan et al. (2008), who use a value of 4 for the *F* statistic for nested GLM models for pre-selection of variables for MaxEnt modelling. Choice of model improvement criterion should guided by experience and by theoretical reasoning. From a theoretical point of view, more reliable and more flexible, statistically based, model improvement criteria should be preferred if available. One of the most important results of this study is that the maximum likelihood explanation of MaxEnt opens for use of standard statistical tools for comparison of nested models, such as the likelihood-ratio test and the *F*-ratio test. Furthermore, the experiments carried out for tuning the *F*-ratio test show that the appropriate degrees of freedom for the residuals in the MaxEnt null model is likely to be $\eta = N - n$, the number of uninformed background observations. However, given the small data sets used for these experiments, the results should be substantiated by experiments on larger data sets.

The PO data sets used for MaxEnt modelling are often strongly biased (e.g., Elith & Graham 2009, Robertson et al. 2010, Wolmarans et al. 2010). From a general statistical point of view, statistical tests with fewer in-built assumptions should then be preferred (e.g., Sokal & Rohlf 1995). This line of reasoning favours the randomisation test over the likelihood-ratio and *F*-ratio tests. However, a disadvantage of the randomisation test is that it, at least so far, cannot be applied to testing of two MaxEnt models of which one contains one or more extra DVs derived from the same EV. The reason for this is that all DVs derived from the same EV are dependent and that a realistic randomisation scheme for the extra DV(s) in the more complex model has not yet been devised. Without such a randomisation scheme, the units subjected to randomisation have to be the EVs themselves and not the single DVs derived from them. Development of randomisation schemes that open for testing of the contribution of single DVs should be encouraged. Likelihood-ratio or *F*-ratio tests therefore have to be used to complete step 3 of the procedure for manual forward stepwise selection of MaxEnt models (Table 4), in which parsimonious sets of DVs are built for each EV.

Like almost all other statistical tests, the likelihood-ratio, the *F*-ratio and randomisation tests for comparison of MaxEnt models assume that the observations are independent replicates drawn from a homogeneous population. It is not clear how this assumption applies to distribution modelling in general and to MaxEnt modelling of distributions in particular; i.e., to *which* population (set of observations) it applies; and *what* is really meant with independence in this context. The result of this study, that the degrees of freedom for the residuals in a MaxEnt model to be used in calculation of the *F* statistic and the associated *p* value is likely to be $N - n$, seemingly indicates that it is the uninformed background observations that should be independent replicates drawn from the population of all possible background cells. In many cases, including the simulated example data sets in this study, all grid cells in the study area are used for modelling: the sample then includes the entire population of uninformed background cells. An alternative way to understand the assumption of independence in the MaxEnt modelling context is by way of the interpretation of MaxEnt probability-ratio output ($\dot{q}$) in continuous explanatory variables space as 'the ratio of the probability of encountering grid cells with environmental characteristics X_l in the subset of presence grid cells to the probability of encountering X_l in the set of all grid cells' or, as expressed by Elith et al. (2011), the 'relative suitability of one place vs. another'. This interpretation of MaxEnt output suggests that it is not bias in presence *or* background grid cells as such that matters, but rather that there is *similar bias in* (samples of) *presence and background* observations (Phillips & Dudík 2008, Elith et al. 2011). Apart from being reasonable from a theoretical point of view, this viewpoint is supported by results of several studies which show that better predictive performance of MaxEnt models on more or less independent evaluation data can be obtained by use of *target-group background*. Target-group background implies that the set of uninformed background cells consists of all grid cells in which any species (or other relevant set of modelled targets) in a taxonomic or ecological entity to which the targeted species belongs is used as background data, rather than the set of all uninformed background cells or a random selection of these cells (Elith & Leathwick 2007, Phillips et al. 2009, Williams et al. 2009, Mateo et al. 2010, Yates et al. 2010). Phillips et al. (2009) show that the performance improvement due to target-group background is largest when there is strong bias in the target-group presence observations.

The extent to which the *p* values in statistical tests for comparison between nested MaxEnt models will be inflated, resulting in Type I error, i.e., falsification of null hypotheses that are actually true (Legendre & Legendre (1998), or otherwise affected, by dissimilar bias in presence and background observations or by spatial autocorrelation in data, requires further study. Such effects are suggested for distribution models by Segurado et al. (2006) and Merckx et al. (2011), among others. Inflation of *p* values by spatial autocorrelation is likely to occur from the perspective that adding uninformed background observations beyond the largest set of spatially non-autocorrelated observations will increase *N*, and hence $\eta = N - n$, without bringing with it the amounts of new, reliable information suggested by the increase in η. This accords with indications in several studies that it is not spatial autocorrelation in the observed presence data and/or the explanatory variables *as such* that is important but spatial autocorrelation remaining in the residuals of the model (Segurado et al. 2006, Dormann et al. 2007, Bini et al. 2009, Franklin et al. 2009, Naimi et al. 2011). The hypothesis that the assumptions of independence and identical distributions primarily applies to the errors (residuals) of MaxEnt models require further study.

The maximum likelihood explanation of MaxEnt opens for calculation of residuals for each grid cell in data sets used for model parameterisation and data sets used for model evaluation, which can then be used to analyse the spatial structure by geostatistical methods as suggested, among others, by Austin (2007) and Dormann (2011). Furthermore, the possibility for incorporating spatial autoregressive terms in MaxEnt models should be further explored. Generalised linear mixed models with autoregressive terms have been shown to improve the performance of other regression-type modelling methods substantially [e.g., Maggini et al. (2006), Diggle &

Ribeiro (2007), Bini et al. (2009), Santika & Hutchinson (2009), Hengl et al. (2009), Carroll et al. (2010); but see Tingley & Herman (2009)].

Potential problems caused by failure of MaxEnt models to fulfill basic assumptions of independence required by standard statistical methods motivate for use of model comparison tests with care. However, even though the *p* values resulting from these tests may turn out to be influenced by bias in data, they are likely to be comparable among models parameterised by use of the same data set. This motivates for parallel use of more than one model improvement criterion (significance level α) when MaxEnt models are built by subset selection methods.

Finally, it should be stressed that the only way to avoid potential problems with lack of comparability of distribution model improvement criteria is to evaluate SPM models by use of P/A data collected independently of data used to parameterise the model. Observations in the evaluation data set should be situated farther apart than the range of the spatial variation of presence observations (cf. Phillips et al. 2009, Veloz 2009, Edvardsen et al. 2011). Evaluation by applying resubstitution and/or data-splitting methods to the PO data set used to parameterise the model does *not* alleviate problems caused by spatial autocorrelation in the data (Araújo et al. 2005, Raes & ter Steege 2007).

PERFORMANCE MEASURES AND THEIR USE FOR QUANTIFYING VARIABLE CONTRIBUTION

The maximum likelihood explanation of MaxEnt provides users with a likelihood-based measure of variation accounted for (VA), *v*, which is analogous to the r^2 of linear models and the deviance of other maximum likelihood modelling methods. However, being based upon the log likelihood of observed presence observations rather than all observations, MaxEnt's measure of variation may differ from these in important properties. The use of log loss to obtain a measure of VA in MaxEnt is not a new idea; this has been common usage since MaxEnt was first made available to the distribution modelling community via the Maxent software in 2004 (e.g., Phillips et al. 2006, Phillips & Dudík 2008). What is new in this paper is the development of log-loss based statistics into methods for comparison of nested MaxEnt models and for quantifying variable contributions to MaxEnt models. Furthermore, the worked examples shed new light on the relationship between the two alternative model performance measures, FTVA (Fraction of Total Variation Accounted for) and AUC (Area Under the receiver operating characteristic Curve).

Results of worked examples show that the scales on which FTVA and AUC are recorded are different scales: even after eventual correction for use with PO data, AUC is recorded on a scale that effectively goes from 0.5 for a random model such as the MaxEnt null model to 1 for a model that perfectly predicts the observed presences and predicts absence in all uninformed background grid cells, such as the MaxEnt saturated model. FTVA, on the other hand, is recorded on a 0–1 scale. Taking this into account, a comparison of the two performance measures for the two example data sets shows that although there seems to be no systematic rank-order inconsistency between the two measures, they do not necessarily follow each other exactly and they are clearly non-linearly related to each other (Fig. 14). Compared to FTVA, the non-parametric AUC measure differentiates strongly between models near the poor-performance end of the scale, i.e., between models that differ little from the null model. Thus, in example data set 1, $\mathrm{AUC}_{\mathrm{corr}} = 0.636$ and FTVA = 0.0422 for the single-variable MaxEnt model for $\boldsymbol{X}_{1,3\mathrm{L}}$ which is found not to be significantly different from the null model both by the *F*-ratio and randomisation tests (*pF* = 0.2551 and *pRand* = 0.1885; Table 8). In this example, the lower 27 % of the effective AUC scale corresponds to the lower 4.2 % of the FTVA scale, i.e., a 'scale utilisation ratio' of ca. 6. Even stronger differentiation between models with poor predictive performance is found in example data set 2: the single-variable MaxEnt model for the L variable derived from the random explanatory variable $\boldsymbol{X}_{2,5L}$ has $\mathrm{AUC}_{\mathrm{corr}}$ = 0.566 (13 %) and FTVA = 0.0090 (0.9 %), corresponding to a 'scale utilisation ratio' of ca. 14. This may indicate a tendency of AUC to emphasise differences between low-performance models more strongly when the number of presence and/or background observations increases. This difference between AUC and FTVA is reflected in measures of variable contribution: measures based upon AUC attribute higher importance to variables with relatively lower explanatory power. These results show that both the non-parametric AUC measure and FTVA can be used to rank MaxEnt models, but also indicate that differences in AUC between models should not be added and subtracted to form variable contribution measures. The two AUC-based measures used in this study ($\mathrm{VC}_{\mathrm{PI}}$ and $\mathrm{VC}_{\mathrm{AUC}}$) strongly emphasise differences in performance between models with low predictive power and, hence, attribute unduly high importance to variables that are likely to be unimportant for the target. One noteworthy result from the worked examples is that among AUC-based measures of variable contribution reported by Maxent software, 'percent contribution' ($\mathrm{VC}_{\mathrm{PC}}$) appears to be much less reliable than the alternative measure, 'permutation importance' ($\mathrm{VC}_{\mathrm{PI}}$). This result, which contrasts the rating of the two measures by Phillips (2011), is exemplified by the single-variable MaxEnt model for the L variable derived from the non-significant EV $\boldsymbol{Z}_{1,2}$, which is not included in any of the manual MaxEnt models. Nevertheless, this variable is attributed a contribution of $\mathrm{VC}_{\mathrm{PI}}$ = 0.683 to the Auto|L model while two other variable contribution measures attribute contributions of 0.012 and 0.045, respectively, to $\boldsymbol{Z}_{1,2}$.

Measures of variation accounted for that are calculated from log loss [expression (29)], i.e., v_t and V_t, and the corresponding measures calculated from regularised log loss [expression (44)], i.e., $\acute{v}_t$ and $\acute{V}_t$, or, what is essentially the same, the 'gain' and 'regularised gain' of Elith et al. (2011) and Phillips (2011), are treated in most MaxEnt modelling studies as if they were commensurable. Although Phillips et al. (2006) and Phillips & Dudík (2008) use two terms, 'log loss' and 'regularised log loss', no distinction seems to be made between them when it comes to use and interpretation [also see Phillips (2011)]. Theoretical reasoning as well as results obtained for the two example data sets in thus study do, however, clearly show that log loss and regularised log loss and model performance statistics and measures of variation calculated from them express different properties of MaxEnt models and are incommensurable (see Tables 10 and 15): while V_t is a likelihood-based measure of variation, statistics calculated from regularised log loss are analogous with penalised likelihood statistics such as AIC and BIC [compare expressions (29) and (45) with expression (44)], expressed on scales without bounds on which the magnitude of differences can hardly be interpreted ecologically. Accordingly, regularised log loss (or 'regularised training gain') should not be used for quantifying relative variable contributions.

Results obtained in the present study suggest that measures of variation calculated from log loss are preferential to AUC for quantifying variable contributions to MaxEnt models. Furthermore, the results suggest that variable contributions should be quantified for EVs represented by a parsimonious set of DVs rather than for the individual DVs. This is illustrated by the individually most important variable in example data set 1, $\boldsymbol{Z}_{1,1}$. When variable contribution is measured by $\mathrm{VC}_{\mathrm{FVA}}$, the relative contribution from $\boldsymbol{Z}_{1,1}$ to the Man+ model, i.e., the model with each variable represented by a parsimonious set consisting of only one DV, is relatively much lower than the variable‘s contribution to the Auto|All models in which this variable is represented by two strongly correlated DVs. Results obtained for four measures of variable contributions in this study show that further research on measures of variable contribution is required to sort out which measures are informative also when the environmental data set contains many, strongly correlated variables.

NEW PERSPECTIVES ON THE GOOD PERFORMANCE OF MAXENT

The default ℓ_1-regularisation procedure is often claimed to be a major reason for MaxEnt's good performance in practical distribution modelling (Hernandez et al. 2006, Phillips et al. 2006, Dudík et al. 2007, Raes & ter Steege 2007, Wisz et al. 2008, Wollan et al. 2008, Tinoco et al. 2009, Elith et al. 2011). Theoretical reasoning and results of worked examples in this paper indicates that MaxEnt performs well despite, and not because of, the ℓ_1-regularisation procedure. Two alternative explanations for MaxEnt's good performance accord with the maximum likelihood explanation of MaxEnt:

1. *Choice of response variable.* The quantity modelled by generative MaxEnt is the probability that one specific presence cell i_0, selected at random from all presence cells, is grid cell *i*. Predictions from MaxEnt models are interpretable as the 'relative suitability of one place vs. another' (Elith et al. 2011). The direct relevance of quantity modelled by MaxEnt for all purposes of distribution modelling and the fact that MaxEnt estimates the response without implicit or explicit assumptions of the prevalence of the modelled target, may contribute to MaxEnt's good performance. It should be noted that the statement in most treatises on MaxEnt that the method is a PO method is not correct; only the generative MaxEnt method is bound to use PO data. MaxEnt shares with other maximum likelihood estimation methods available for distribution modelling with PO data, e.g., GLM and GAM, the property that uninformed background observations are treated as pseudo-absence observations. This is evident from the fact that the reference model for with which all other MaxEnt models are compared, the saturated model, predicts absence ($\pi = 0$) in all uninformed background cells and presence ($\pi = \frac{1}{n}$) in all observed presence cells [expression (15)].
2. *Flexibility and ecological realism of the fitted functional relationship*. The worked examples show that the Gibbs function fitted by MaxEnt has the flexibility needed to model overall ecological response curves with a large variety of realistic shapes: linear, plateau-shaped, symmetric and skewed, unimodal or truncated unimodal. However, this flexibility with respect to response-curve shapes is not an inherent property of the MaxEnt method, but a property that *may* arise if the range of transformation functions used to derive DVs from EVs is appropriate. The input to MaxEnt is single (derived) variables, which are combined to more or less complex ecological response models. MaxEnt also handles interactions between variables. MaxEnt thus opens for fitting the entire range of realistic functional relationships between response and explanatory variables and thus combines attributes of GLM and GAM. The similarity between MaxEnt on one hand and GLM and GAM on the other with respect to flexibility in model fitting is also pointed out by Phillips et al. (2006), Suárez-Seoane et al. (2008), Willems & Hill (2009) and Elith et al. (2011).

CONCLUDING REMARKS, RECOMMENDATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

MaxEnt became a state-of-the art method for distribution modelling in less than five years after the method was made available to a broad audience of distribution modellers. MaxEnt's success is due to good documented performance in practical distribution modelling and easy access via the free, user-friendly Maxent software. In this paper I combine the conceptual framework of the gradient analytic perspective on distribution modelling (Halvorsen 2012) with maximum likelihood estimation into a new explanation of MaxEnt. This theoretical platform, supported by simple worked examples, opens several possibilities for improvement of the current standard practice for distribution modelling by MaxEnt. These improvements are partly methodological, such as changes of options and settings or implementation of tools currently in use with other methods, such as the likelihood-ratio and *F*-ratio tests, partly practical, such as suggestions for new tools that can be operationalised for use directly in Maxent software, or indirectly, e.g., as R tools that work together with Maxent software (Hijmans & Elith 2011, Phillips 2011).

The potentially most important methodological improvements suggested in the present study, and the practical tools needed to implement them, are:

1. Flexible, interactive tools to assist the process by which derived variables are obtained from explanatory variables by the transformation step (Step 5,ii in the 12-step DM process), including: (i) graphical tools such as the frequency-of-observed-presence plots to assist choice of derived variable type; (ii) a broader range of flexible functions for transformation of explanatory variables, including monotonous functions, deviation-type functions, and complex spline functions; and (iii) V_t- knot graphs to guide transformation of explanatory variables into derived variables of the spline types.
2. A comprehensive, flexible, interactive toolbox that allows the user to combine (i) model selection methods; (ii) methods and approaches for internal model performance assessment, and (iii) model improvement criteria, that opens for integration of independent presence/absence data into the modelling process, for external model performance assessment, for model calibration, and for model evaluation. Model selection tools should include the full range of manual and automated procedures for (a) subset selection, including forward and forward–backward, for small sets of explanatory variables perhaps also backward, selection; and (b) shrinkage methods (see Dudík et al. 2007). Facilities for interactive manual selection should be available at each step in the modelling process, thus allowing for manual construction of parsimonious sets of derived variables for each explanatory variable as well as manual construction of multi-variable MaxEnt models. Methods and approaches for internal model performance assessment should include likelihood-ratio and *F*-ratio tests, randomisation tests, and AUC-based methods, with model improvement criteria set by the user.
3. New output formats; (i) the probability-ratio output format $\dot{q}$, which expresses the 'relative suitability of one place vs. another', should be available for cases by which no presence/absence data are available for calibrating the output to a probability-of-presence scale; and (ii) the probability-of-presence output format $\breve{q}$, which expresses the predicted probability of presence in a site, should be available for cases by which independent presence/absence data are available.
4. Options for discriminative use of MaxEnt, i.e., MaxEnt modelling by use of presence/absence data.

Many of these tools are accessible today; some are implemented in the Maxent software (Elith et al. 2011, Phillips 2011) and some are available in R tools more or less well integrated with Maxent software (Elith et al. 2011, Hijmans & Elith 2011, Phillips 2011). Most of the proposed tools do, however, require extensive scripting or programming to be made accessible. Exploration of the proposed new options, settings and tools will depend on accessibility, e.g., implementation in user-friendly software such as Maxent software and/or in R as a 'MaxEnt for R' library, and/or in other new software. An outline of a consensus MaxEnt practice, applicable for spatial prediction modelling (SPM), general-purpose ecological response modelling, and most projective distribution modelling (PPM) purposes (Halvorsen 2012), that emerge from theoretical reasoning, examples and discussion throughout this paper, is given in Table 17.

Several research needs have been identified, among which the most important are considered to be:

1. Further exploration of statistical properties of the MaxEnt method and associated tools from a maximum likelihood modelling perspective, e.g.: (i) effects of spatial autocorrelation and other aspects of non-independence of response and exploratory variables for statistical inference about model performance; (ii) the possibility for incorporating spatial autoregressive terms in MaxEnt models; (iii) determination of appropriate degrees of freedom for sets of variables derived from one explanatory variable by transformation; (iv) investigations into the statistical properties of the measure of variation accounted for in MaxEnt, which is based upon log loss, and its relationship to the deviance; (v) further development of existing methods and approaches for internal model performance assessment, and development of new measures of model performance; and (vi) development of improved measures of variable contribution to models.
2. Research to find best strategies for transformation of explanatory variables into derived variables and for construction of parsimonious sets of derived variables that account for unimodality, skewness and/or platy- or leptokurtosis in frequency-of-presence curves for targets subjected to distribution modelling by MaxEnt.
3. Extensive comparative tests of the predictive performance of MaxEnt models over the entire range of realistic choices of options and settings, including model selection methods, internal model performance assessment, and model improvement criteria, in search for patterns with general applicability. A particularly important question is if one, generally applicable, procedure for parameterisation of MaxEnt models can be found. The importance of using independent presence/absence data for evaluation of distribution models for the SPM purpose, including their options and settings, cannot be too strongly emphasised.

A summary of research needs is given in Table 17.

I hope this paper will contribute to a better understanding of what goes on in 'the MaxEnt black box' and stimulate research on the still many obscure issues in MaxEnt methodology. I also hope that the results have demonstrated clearly that the general recommendation not to trust automated procedures blindly (Økland 2007) also applies to MaxEnt. I hope this paper will stimulate further development of user-friendly tools for MaxEnt modelling, which may in turn assist the search for generally robust principles for modelling by MaxEnt. If models with near-optimal predictive performance on a given data set can be guaranteed, MaxEnt will be an even better tool for practical distribution modelling, e.g., for conservation purposes. Finally, I hope this paper will stimulate comparative studies of MaxEnt options and settings, in continuous search for improved MaxEnt modelling practices.

Table 17. Distribution modelling by MaxEnt: summary of recommendations and research needs, structured according to the 12-step distribution modelling process of Halvorsen (2012); see Fig. 1. Support, empirical or theoretical, is indicated on a five-point scale: none, weak, some, good, strong.

Step	Issue	Recommendation	Support	Research needs
(5,ii)	Transformation of explanatory variables into derived variables			* development of estimators of appropriate degrees of freedom for sets of variables derived from one explanatory variable by transformation
	Choice of type of derived variables	Open for inclusion of all types of derived variables, perhaps except variables of the T type, and use frequency-of-observed-presence plots to judge the appropriateness of each type	strong	* assessment of the general usefulness of DVs of different types for MaxEnt modelling
	Choice of M-type variables	Open for inclusion of several M-type variables obtained by different nonlinear transformations, e.g., the by the zero-skewness transformation	good	* assessment of the importance of nonlinear transformations into M-type variables for model performance
	Choice of D-type variables	Open for D-type variables of the general form given by expression (102) in Table 2, i.e., that express deviation from the estimated optimum of the modelled target	strong	* optimalisation of the parameter *a* in expression (102)
	Choice of spline variables	Use the Vt-knot graph to select one or a few FROM the unlimited pool of potential spline variables	strong	
	Use of X-type variables			* comparative tests of MaxEnt model performance with X-type variables, e.g., obtained by HOF modelling, vs other sets of DVs
	Use of interaction variables			* assessment of the ability of interaction variables to improve MaxEnt model performance
(7,ii)	Model specification			* explore the possibility for incorporating spatial autoregressive terms in MaxEnt models
	Use of ‘target-group background’	Use of target-group background observations is recommended, unless there are good reasons for choosing another option	some	* comparative studies of MaxEnt models with and without use of ‘target-group background’
(8,i)	Model selection			* by extensive comparative studies, to establish a ‘consensus MaxEnt practice’ for SPM purposes * explore context-driven tuning of MaxEnt in search for a ‘best specific MaxEnt practice’

Table 17 (Continued).

Step	Issue	Recommendation	Support	Research needs
	Choice of model selection strategy	Subset selection methods such as the manual forward stepwise selection procedure outlined in Table 4, should be preferred to the 'standard Maxent procedure' with ℓ_1-regularisation	good	* comprehensive comparison of model selection strategies, and of settings and options for each strategy
	Choice of appropriate level of model complexity	Should be determined after due consideration of modelling purpose; ERM and PPM require simpler models than SPM	strong	* comprehensive comparison of the predictive capability of models with different complexities
(8,ii) Methods and approaches for internal model performance assessment, including choice of model improvement criterion				
	Appropriate degree of regularisation	Should be determined after due consideration of modelling purpose and properties of the modelled target. In the absence of other indications, more strict model improvement criteria (i.e., higher value of the regularisation parameter λ) than normally applied is recommended (e.g., much lower α levels than 0.05 in likelihood-ratio, and *F*-ratio tests)	good	* comprehensive comparison of the predictive capability of models obtained by the same model selection strategy, but differing with respect to strictness of the model improvement criterion * assessment of the extent to which the appropriate degree of regularisation can be decided from properties of the modelled target
	Choice of method for internal model performance assessment	Statistical tests, such as the likelihood-ratio and the *F*-ratio tests are recommended for variable selection	some	* development of randomisation schemes that open for testing of single derived variables * comparative studies of statistical tests for comparison of nested MaxEnt models * tests on large simulated data sets to confirm (or reject) the proposal in this study that the appropriate degrees of freedom for the residuals in the *F*-ratio test is $\eta = N - n$ * assessments of the effects of spatial autocorrelation or other biases in data on the reliability of model comparisons by statistical tests
(8,iii) Quantifying variable contribution				* better understanding of the statistical properties of the measure of variation accounted for in MaxEnt and its relationship to r^2 and deviance of GLM and GAM
	Use of penalised performance statistics to quantify variable contributions	Use of penalised performance statistics to quantify variable contributions should be avoided	strong	

Table 17 (Continued).

Step	Issue	Recommendation	Support	Research needs
	Choice of variable contribution measure	Variable contribution measures calculated from (unregularised) log loss, v_t and V_t, should be preferred to AUC-based measures	good	* comparative studies of variable contribution measures * development of improved measures of variable contribution, e.g., measures that account properly for the contribution of environmental variables represented by several derived variables
(8,iv) Output formats				
	Choice of output format	Without access to independent P/A data, the probability-ratio output format $\dot{q}$ is recommended unless there are good reasons to choose another format	good	
(10 & 11) Model calibration and model evaluation				
	Importance of independent P/A data	Access to P/A data, collected independently of the PO data used to parameterise the model, and model calibration and evaluation based upon such data, is essential for all models that shall serve the SPM purpose	strong	
		With access to independent P/A data, calibration to a probability-of-presence output format is recommended	strong	

ACKNOWLEDGEMENTS

This study has benefitted strongly from long-term exchange of ideas with, and constructive criticism from, members and affiliates of the geo-ecology research group (formerly 'Identification and modelling of biodiversity') at the Natural History Museum, University of Oslo.. In particular, I thank Vegar Bakkestuen for continuous support; Inger Auestad, Trine Bekkby, Anders Bryn, Lars Erikstad, Knut Anders Hovstad, Eli Rinde and Jogeir N. Stokland and former and present PhD students John W. Dirksen, Anette Edvardsen, Sabrina Mazzoni, Thijs van Son, Bente Støa, Anders K. Wollan for comments on and suggestions for improvement of earlier versions of this manuscript. Furthermore, all are thanked for contributions to the empirical distribution modelling studies, published or in preparation, which form the basis for most insights reported in this paper. Steven Phillips is thanked for practical assistance with Maxent software in an early phase of this study. Special thanks go to Einar Heegaard whose detailed, critical comments on an earlier version of the manuscript considerably improved the final result.

REFERENCES

* = included in mini-review of studies using MaxEnt

Akaike, H. 1973. Information theory as an extension of the maximum likelihood principle. – In: Petrov, B.N. & Csaki, F. (eds), Second international symposium on information theory, Akademiai Kiado, Budapest, pp. 267–281.

*Anderson, R.P. & Gonzalez, I.J. 2011. Species-specific tuning increases robustness to sampling bias in models of species distributions: an implementation with Maxent. – Ecol. Modelling 222: 2796-2811.

*Anderson, R.P. & Raza, A. 2010. The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus *Nephelomys*) in Venezuela. – J. Biogeogr. 37: 1378-1393.

Anonymous 2010. R version 2.11 for Windows. – The R foundation for statistical computing (http://cran.r-project.org)

*Aranda, S.C. & Lobo, J.M. 2011. How well does presence-only-based species distribution modelling predict assemblage diversity? A case study of the Tenerife flora. – Ecography 34: 31-38.

Araújo, M.B. & Guisan, A. 2006. Five (or so) challenges for species distribution modelling. – J. Biogeogr. 33: 1677-1688.

Araújo, M.B., Pearson, R.G., Thuiller, W. & Erhard, M. 2005. Validation of species-climate impact models under climate change. – Global Change Biol. 11: 1504-1513.

*Auestad, I., Halvorsen, R., Bakkestuen, V. & Erikstad, L. 2011. Utbredelsesmodellering av fremmede invaderende karplanter langs veg. – Dir. Naturforv. Utredn. 2011: 2: 1-30.

Austin, M.P. 1999. A silent clash of paradigms: some inconsistencies in community ecology. – Oikos 86: 170-178.

Austin, M.P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. – Ecol. Modelling 157: 101-118.

Austin, M. 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. – Ecol. Modelling 200: 1-19.

Bakkestuen, V., Aarrestad, P.A., Stabbetorp, O.E., Erikstad, L. & Eilertsen, O. 2010. Vegetation composition, gradients and environment relationships of birch forest in six reference areas in Norway. – Sommerfeltia 33: 1-226.

Barbosa, A.M. 2009. Transferability of environmental favourability models in geographic space: the case of the Iberian desman (*Galemys pyrenaicus*) in Portugal and Spain. – Ecol. Modelling 220: 747-754.

Barry, S. & Elith, J. 2006. Error and uncertainty in habitat models. – J. appl. Ecol. 43: 413-423.

*Bartel, R.A. & Sexton, J.O. 2009. Monitoring habitat dynamics for rare and endangered species using satellite images and niche-based models. – Ecography 32: 888-896.

*Bedia, J., Busqué, J. & Gutiérrez, J.M. 2011. Predicting plant species distribution across an alpine rangeland in northern Spain: a comparison of probabilistic methods. – Appl. Veg. Sci. 14: 415-432.

Berger, A.L., Della Pietra, S.A. & Della Pietra, V.J. 1996. A maximum entropy approach to natural language processing. – Comput. Linguist. 22: 39-71.

Bini, L.M., Diniz-Filho, A.F., Rangel, T.F.L.V.B., Akre, T.S.B., Albaladejo, R.G., Albuquerque, F.S., Aparicio, A., Araújo, M.B., Baselga, A., Beck, J., Bellocq, M.I., Böhning-Gaese, K., Borges, P.A.V., Castro-Parga, I., Chey, V.K., Chown, S.L., de Marco, P.J., Dobkin, D.S., Ferrer-Castán, D., Field, R., Filloy, J., Fleishman, E., Gómez, J.F., Hortal, J., Iverson, J.B., Kerr, J.T., Kissling, W.D., Kitching, I.J., León-Cortés, J.L., Lobo, J.M., Montoya, D., Morales-Castilla, I., Moreno, J.C., Oberdorff, T., Olalla-Tárraga, M.Á., Pausas, J.G., Qian, H., Rahbek, C., Rodríguez, M.Á., Rueda, M., Ruggiero, A., Sackmann, P., Sanders, N.J., Terribile, L.C., Vetaas, O.R. & Hawkins, B.A. 2009. Coefficient shifts in geographical ecology: an empirical evaluation of spatial and non-spatial regression. – Ecography 32: 193-204.

*Bradley, B.A., Wilcove, D.S. & Oppenheimer, M. 2010. Climate change increases risk of plant invasion in the Eastern United States. – Biol. Invasions 12: 1855-1872.

*Braunisch, V. & Suchant, R. 2010. Predicting species distributions based on incomplete survey data: the trade-off between precision and scale. – Ecography 33: 826-840.

Brown, K.A., Spector, S. & Wu, W. 2008. Multi-scale analysis of species introductions: combining landscape and demographic models to improve management decisions about non-native species. – J. appl. Ecol. 45: 1639-1648.

*Buermann, W., Saatchi, S., Smith, T.B., Zutta, B.R., Chaves, J.A., Milá, B. & Graham, C.H. 2008. Predicting species distributions across the Amazonian and Andean regions using remote sensing data. – J. Biogeogr. 35: 1160-1176.

Burnham, K.P. & Anderson, D.R. 2002. Model selection and multimodel inferences: a practical information-theoretic approach, ed. 2. – Springer, New York.

*Carnaval, A.C. & Moritz, C. 2008. Historical climate modelling predicts patterns of current biodiversity in the Brazilian Atlantic forest. – J. Biogeogr. 35: 1187-1201.

*Carroll, C., Johnson, D.S., Dunk, J.R. & Zielinski, W.J. 2010. Hierarchical Bayesian spatial models for multispecies conservation planning and monitoring. – Conserv. Biol. 24: 1538-1548.

*Cordellier, M. & Pfenninger, M. 2009. Inferring the past to predict the future: climate modelling predictions and phylogeography for the freshwater gastropod *Radix balthica* (Pulmonata Basommatophora). – Molec. Ecol. 18: 534-544.

*Costa, G.C., Nogueira, C., Machado, R.B. & Colli, G.R. 2010. Sampling bias and the use of ecological niche modeling in conservation planning: a field evaluation in a biodiversity hotspot. – Biodiv. Conserv. 19: 883-899.

Crawley, M.J. 2007. The R book. – Wiley, Chichester.

*Cunningham, H.R., Rissler, L.J. & Apodaca, J.J. 2009. Competition at the boundary in the slimy salamander: using reciprocal transplants for studies on the role of biotic interactions in spatial distributions. – J. Anim. Ecol. 78: 52-62.

Danz, N.P., Reich, P.B., Frelich, L.E. & Niemi, G.J. 2011. Vegetation controls vary across space and spatial scale in a historic grassland-forest biome boundary. – Ecography 34: 402-414.

De'ath, G. 2007. Boosted trees for ecological modeling and prediction. – Ecology 88: 243-251.

Della Pietra, S., Della Pietra, V. & Lafferty, J. 1997. Inducing features of random fields. – IEEE Trans. Pattern Anal. Mach. Intell. 19: 1-13.

DeLong, E.R., DeLong, D.M. & Clarke-Pearson, D.L. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a non-parametric approach – Biometrics 44: 837-845.

*DeMatteo, K.E. & Loiselle, B.A. 2008. New data on the status and distribution of the bush dog (Speothos venaticus): evaluating its quality of protection and directing research efforts. – Biol. Conserv. 141: 2494-2505.

*Diniz-Filho, J.A.F., Bini, L.M., Rangel, T.F., Loyola, R.D., Hof, C., Nogués-Bravo, D. & Araújo, M.B. 2009. Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. – Ecography 32: 897-906.

Dobrowski, S.Z., Safford, H., D., Cheng, Y.B. & Ustin, S.L. 2008. Mapping mountain vegetation using species distribution modeling, image-based texture analysis, and object-based classification. – Appl. Veg. Sci. 11: 499-508.

Dormann, C. 2011. Modelling species' distributions. – In: Jopp, F., Reuter, H. & Breckling, B. (eds), Modelling complex ecological dynamics: an introduction into ecological modelling, Springer, Berlin, pp. 179-196.

Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Kissling, W.D., Kühn, I., Ohlemüller, R., Peres-Neto, P.R., Reineking, B., Schröder, B., Schurr, F.M. & Wilson, R. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. – Ecography 30: 609-628.

Dubuis, A., Pottier, J., Rion, V., Pellissier, L., Theurillat, J.-P. & Guisan, A. 2011. Predicting spatial patterns of plant species richness: a comparison of direct macroecological and species stacking modelling approaches. – Divers. Distrib. 17: 1122-1131.

Dudík, M. & Phillips, S.J. 2009. Generative and discriminative learning with unknown labeling bias. – Adv. neural Inf. Process. Syst. 21: 401-408.

Dudík, M., Phillips, S.J. & Schapire, R.E. 2007. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. – J. Machine Learning Res. 8: 1217-1260.

*Echarri, F., Tambussi, C. & Hospitaleche, C.A. 2009. Predicting the distribution of the crested tinamous, *Eudromia* spp. (Aves, Tinamiformes). – J. Ornithol. 150: 75-84.

*Edrén, S.M.C., Wisz, M.S., Teilmann, J., Dietz, R. & Söderkvist, J. 2010. Modelling spatial patterns in harbour porpoise satellite telemetry data using maximum entropy. – Ecography 33: 698-708.

*Edvardsen, A., Bakkestuen, V. & Halvorsen, R. 2011. A fine-grained spatial prediction model for the red-listed vascular plant *Scorzonera humilis*. – Nord. J. Bot. 29: 495-504.

Edwards, T.C.J., Cutler, D.R., Zimmermann, N.E., Geiser, L. & Alegria, J. 2005. Model-based stratifications for enhancing the detection of rare ecological events. – Ecology 86: 1081-1090.

Edwards, T.C.J., Cutler, D.R., Zimmermann, N.E., Geiser, L. & Moisen, G.G. 2006. Effects of sample survey design on the accuracy of classification tree models in species distribution models. – Ecol. Modelling 199: 132-141.

Elith, J. & Graham, C.H. 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. – Ecography 32: 66-77.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, 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., Overton, J.M.M., 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.

*Elith, J., Kearney, M. & Phillips, S. 2010. The art of modelling range-shifting species. – Methods Ecol. Evol. 1: 330-342.

Elith, J. & Leathwick, J.R. 2009. Species distribution models: ecological explanation and prediction across space and time. – A. Rev. Ecol. Evol. Syst. 40: 677-697.

Elith, J., Leathwich, J.R. & Hastie, T. 2008. A working guide to boosted regression trees. – J. Anim. Ecol. 77: 802-813.

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.

*Feeley, K.J. & Silman, M.R. 2011. Keep collecting: accurate species distribution modelling requires more collections than previously thought. – Divers. Distrib. 17: 1132-1140.

*Ficetola, G.F., Thuiller, W. & Miaud, C. 2007. Prediction and validation of the potential global distribution of a problematic invasive alien species. – Divers. Distrib. 13: 476-485.

Fielding, A.H. & Bell, J.E. 1997. A review of methods for the assessment of prediction errors in conservation presence-absence models. – Environm. Conserv. 24: 38-49.

*Fitzpatrick, M.C., Gove, A.D., Sanders, N.J. & Dunn, R.R. 2008. Climate change, plant migration, and range collapse in a global biodiversity hotspot: the *Banksia* (Proteaceae) of Western Australia. – Global Change Biol. 14: 1337-1352.

Franklin, J. 2009. Mapping species distributions: spatial inference and prediction. – Cambridge University Press, Cambridge.

Franklin, J. 2010. Moving beyond static species distribution models in support of conservation biogeography. – Divers. Distrib. 16: 321-330.

Franklin, J., Wejnert, K.E., Hathaway, S.A., Rochester, C.J. & Fisher, R.N. 2009. Effect of species rarity on the accuracy of species distribution models for reptiles and amphibians in southern California. – Divers. Distrib. 15: 167-177.

Friedman, J.H. 1991. Multivariate adaptive regression splines. – Ann. Statist. 19: 1-67.

*Gaikwad, J., Wilson, P.D. & Ranganathan, S. 2011. Ecological niche modeling of customary medicinal plant species used by Australian aborigines to identify species-rich and culturally valuable areas for conservation. – Ecol. Modelling 222: 3437-3443.

*Gastón, A. & García-Viñas, J.I. 2011. Modelling species distributions with penalised logistic regressions: a comparison with maximum entropy models. – Ecol. Modelling 222: 2037-2041.

Gellrich, M. & Zimmermann, N.E. 2007. Investigating the regional-scale pattern of agricultural land abandonment in the Swiss mountains: a spatial statistical modelling approach. – Landsc. Urban Planning 79: 65-76.

*Gibson, L., Barrett, B. & Burbridge, A. 2007. Dealing with uncertain absences in habitat modelling: a case study of a rare ground-dwelling parrot. – Divers. Distrib. 13: 704-713.

Gini, C. 1912. Variabilità e mutabilità – Bologna, Cuppini.

*Giovanelli, J.G.R., Haddad, C.F.B. & Alexandrino, J. 2008. Predicting the potential distribution of the alien invasive American bullfrog (*Lithobates catesbeianus*) in Brazil. – Biol. Invasions 10: 585-590.

*Gormley, A.M., Forsyth, D.M., Griffioen, P., Lindeman, M., Ramsey, D.S.L., Scroggie, M.P. & Woodford, L. 2011. Using presence-only and presence-absence data to estimate the current and potential distributions of established invasive species. – J. appl. Ecol. 48: 25-34.

*Graham, C.H. & Hijmans, R.J. 2006. A comparison of methods for mapping species ranges and species richness. – Global Ecol. Biogeogr. 15: 578-587.

*Graham, C.H., VanDerWal, J., Phillips, S.J., Moritz, C. & Williams, S.E. 2010. Dynamic refugia and species persistence: tracking spatial shifts in habitat through time. – Ecography 33: 1062-1069.

Guisan, A., Broennimann, O., Engler, R., Vust, M., Yoccoz, N.G., Lehmann, A. & Zimmermann, N.E. 2006. Using niche-based models to improve the sampling of rare species. – Conserv. Biol. 20: 501-511.

Guisan, A., Graham, C.H., Elith, J., Huettmann, F. & Group, N.S.D.M. 2007. Sensitivity of predictive species distribution models to change in grain size. – Divers. Distrib. 13: 332-340.

Guisan, A. & Zimmermann, N.E. 2000. Predictive habitat distribution models in ecology. – Ecol. Modelling 135: 147-186.

Halvorsen, R. 2012. A gradient analytic perspective on distribution modelling. – Sommerfeltia, submitted manuscript.

Hanley, J.A. & McNeil, B.J. 1982. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. – Radiology 143: 29-36.

Hastie, T., Tibshirani, R. & Friedman, J. 2009. The elements of statistical learning, ed. 2. – Springer, New York.

Hengl, T., Sierdsema, H., Radović, A. & Dilo, A. 2009. Spatial prediction of species' distributions from occurrence-only records: combining point pattern analysis, ENFA and regression-kriging. – Ecol. Modelling 220: 3499-3511.

*Hernandez, P.A., Graham, C.H., Master, L.L. & Albert, D.L. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. – Ecography 29: 773-785.

Hijmans, R.J. & Elith, J. 2011. Species distribution modelling with R. – http://cran.r-project.org/web/packages/dismo/vignettes/sdm.pdf, The R foundation for statistical computing.

*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 Biol. 12: 2272-2281.

Hirzel, A.H., Le Lay, G., Helfer, V., Randin, C. & Guisan, A. 2006. Evaluating the ability of habitat suitability models to predict species presences. – Ecol. Modelling 199: 142-152.

Hjort, J. & Marmion, M. 2009. Periglacial distribution modelling with a boosting method. – Permafr. periglac. Proc 20: 15-25.

*Hoffman, J.D., Aguilar-Amuchastegui, N. & Tyre, A.J. 2010. Use of simulated data from a process-based habitat model to evaluate methods for predicting species occurrence. – Ecography 33: 656-666.

Hortal, J., Jiménez-Valverde, A., Gómez, J.F., Lobo, J.M. & Baselga, A. 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. – Oikos 117: 847-858.

Huisman, J., Olff, H. & Fresco, L.F.M. 1993. A hierarchical set of models for species response analysis. – J. Veg. Sci. 4: 37-46.

Jaynes, E.T. 1957. Information theory and statistical mechanics. – Phys. Rev. 106: 620-630.

Jaynes, E.T. 1957. Information theory and statistical mechanics. II. – Phys. Rev. 108: 171-190.

Jaynes, E.T. 2003. Probability theory: the logic of science. – Cambridge University Press, Cambridge.

Jiménez-Valverde, A., Lobo, J. & Hortal, J. 2008. Not as good as they seem: the importance of concepts in species distribution modelling. – Divers. Distrib. 14: 885-890.

Kadmon, R., Farber, O. & Danin, A. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. – Ecol. Appl. 14: 401-413.

*Kharouba, H.M., Algar, A.C. & Kerr, J.T. 2009. Historically calibrated predictions of butterfly species' range shift using global change as a pseudo-experiment. – Ecology 90: 2213-2222.

*Ko, C.Y., Root, T.L. & Lee, P.F. 2011. Movement distances enhance validity of predictive models. – Ecol. Modelling 222: 947-954.

Kodric-Brown, A. & Brown, J.H. 1993. Incomplete data sets in community ecology and biogeography: a cautionary tale. – Ecol. Appl. 3: 736–742.

Kullback, S. 1959. Information theory and statistics. – New York, Wiley.

*Lahoz-Monfort, J.J., Guillera-Arroita, G., Milner-Gulland, E.J., Young, R.P. & Nicholson, E. 2010. Satellite imagery as a single source of predictor variables for habitat suitability modelling: how Landsat can inform the conservation of a critically endangered lemur. – J. appl. Ecol. 47: 1094-1102.

*Lamb, J.M., Ralph, T.M.C., Goodman, S.M., Bogdanowicz, W.,, Fahr, J., Gajewska, M., Bates, P.J.J., Eger, J., Benda, P. & & Taylor, P.J. 2008. Phylogeography and predicted distribution of African-Arabian and Malagasy populations of giant mastiff bats, *Otomops* spp. (Chiroptera: Molossidae). – Acta chiropterol. 10: 21-40.

Leathwick, J.R., Elith, J. & Hastie, T. 2006. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. – Ecol. Modelling 199: 188-196.

Legendre, P. & Legendre, L. 1998. Numerical ecology, ed. 2. – Elsevier, Amsterdam.

Lobo, J.M. 2008. More complex distribution models or more representative data? – Biodiv. Inform. 5: 14-19.

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 Ecol. Biogeogr. 17: 145-151.

*Loiselle, B.A., Jørgensen, P.M., Consiglio, T., Jiménez, I., Blake, J.G., Lohmann, L.G. & Montiel, O.M. 2008. Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? – J. Biogeogr. 35: 105-116.

*Lozier, J.D., Aniello, P. & Hickerson, M.J. 2009. Predicting the distribution of Sasquatch in western North America: anything goes with ecological niche modelling. – J. Biogeogr. 36: 1623-1627.

Luoto, M., Pöyry, J., Heikkinen, R.K. & Saarinen, K. 2005. Uncertainty of bioclimatic envelope models based on the geographical distribution of species. – Global Ecol. Biogeogr. 14: 575-584.

McCarthy, K.P., Fletcher, R.J.J., Rota, C.T. & Hutto, R.L. 2011. Predicting species distributions from samples collected along roadsides. – Conserv. Biol. 26: 68-77.

Maggini, R., Lehmann, A., Zimmermann, N.E. & Guisan, A. 2006. Improving generalized regression analysis for the spatial prediction of forest communities. – J. Biogeogr. 33: 1729-1749.

*Marini, M.Á., Barbet-Massin, M., Lopes, L. & Jiguet, F. 2010. Predicting the occurrence of rare Brazilian birds with species distribution models. – J. Ornithol. 151: 857-866.

*Marino, J., Bennett, M., Cossios, D., Iriarte, A., Lucherini, M., Pliscoff, P., Sillero-Zubiri, C., Villalba, L. & Walker, S. 2011. Bioclimatic constraints to Andean cat distribution: a modelling application for rare species. – Divers. Distrib. 17: 311-322.

Marmion, M., Luoto, M., Heikkinen, R.K. & Thuiller, W. 2009a. The performance of state-of-the-art modelling techniques depends on geographical distribution of species. – Ecol. Modelling 220: 3512-3520.

*Mateo, R.G., Croat, T.B., Felicísimo, Á.M. & Muñoz, J. 2010. Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections. – Divers. Distrib. 16: 84-94.

*Merckx, B., Steyaert, M., Vanreusel, A., Vincx, M. & Vanaverbeke, J. 2011. Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. – Ecol. Modelling 222: 588-597.

Metz, C.E. 1978. Basic principles of ROC curve analysis. – Semin. nucl. Med. 8: 283-298.

Michel, P., Overton, J.M., Mason, N.W.H., Hurst, J.M. & Lee, W.G. 2011. Species-environment relationships of mosses in New Zealand indigenous forest and shrubland ecosystems. – Pl. Ecol. 212: 353-367.

Minchin, P.R. 1989. Montane vegetation of the Mt. Field massif, Tasmania: a test of some hypotheses about properties of community patterns. – Vegetatio 83: 97-110.

*Monterroso, P., Brito, J.C., Ferreras, P. & Alves, P.C. 2009. Spatial ecology of the European wildcat in a Mediterranean ecosystem: dealing with small radio-tracking datasets in species conservation. – J. Zool. 279: 27-35.

Mouton, A.M., de Baets, B. & Goethals, P.L.M. 2010. Ecological relevance of performance criteria for species distribution models. – Ecol. Modelling 221: 1995-2002.

Murphy, A.H. & Winkler, R.L. 1987. A general framework for forecast verification. – Mon. Weather Rev. 115: 1330-1338.

*Murray-Smith, C., Brummitt, N.A., Oliveira-Filho, A.T., Bachman, S., Moat, J., Lughadha, E.M.N. & Lucas, E.J. 2009. Plant diversity hotspots in the Atlantic coastal forests of Brazil. – Conserv. Biol. 23: 151-163.

Myers, R.H., Montgomery, D.C. & Vining, G.G. 2002. Generalized linear models with applications in engineering and the sciences. – Wiley, New York.

*Niamir, A., Skidmore, A.K., Toxopeus, A.G., Muñoz, A.R. & Real, R. 2011. Finessing atlas data for species distribution models. – Divers. Distrib. 17: 1173-1185.

Nóbrega, C.C. & de Marco, P.J. 2011. Unprotecting the rare species: a niche-based gap analysis for odonates in a core Cerrado area. – Divers. Distrib. 17: 491-505.

Økland, R.H. 1990a. Vegetation ecology: theory, methods and applications with reference to Fennoscandia. – Sommerfeltia Suppl. 1: 1-233.

Økland, R.H. 1992. Studies in SE Fennoscandian mires: relevance to ecological theory. – J. Veg. Sci. 3: 279-284.

Økland, R.H. 2007. Wise use of statistical tools in ecological field studies. – Folia geobot. 42: 123-140.

Økland, R.H., Rydgren, K. & Økland, T. 2003. Plant species composition of boreal spruce swamp forests: closed doors and windows of opportunity. – Ecology 84: 1909-1919.

Økland, R.H., Økland, T. & Rydgren, K. 2001. Vegetation-environment relationships of boreal spruce swamp forests in Østmarka Nature Reserve, SE Norway. – Sommerfeltia 29: 1-190.

Oksanen, J. & Minchin, P.R. 2002. Continuum theory revisited: what shape are species responses along ecological gradients? – Ecol. Modelling 157: 119-129.

*Parisien, M.A. & Moritz, M.A. 2009. Environmental controls on the distribution of wildfire at multiple spatial scales. – Ecol. Monogr. 79: 127-154.

*Parolo, G., Rossi, G. & Ferrarini, A. 2008. Toward improved species niche modelling: Arnica montana in the Alps as a case study. – J. appl. Ecol. 45: 1410-1418.

Pearce, J.L. & Boyce, M.S. 2006. Modelling distribution and abundance with presence-only data. – J. appl. Ecol. 43: 405-412.

Pearce, J. & Ferrier, S. 2000a. An evaluation of alternative algorithms for fitting species distribution models using logistic regression. – Ecol. Modelling 128: 127-147.

Pearce, J.L. & Ferrier, S. 2000b. Evaluating the predictive performance of habitat models developed using logistic regression. – Ecol. Modelling 133: 225-245.

*Pearson, R.G., Raxworthy, C.J., Nakamura, M. & Peterson, A.T. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. – J. Biogeogr. 34: 102-117.

Peterson, A.T., Papes, M. & Eaton, M. 2007. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. – Ecography 30: 550-560.

Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E., Nakamura, M. & Araújo, M.B. 2011. Ecological niches and geographic distributions. – Monogr. Pop. Biol. 49: 1-314.

Phillips, S.J. 2011. A brief tutorial on Maxent. – AT&T Research, Princeton, NJ.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Modelling 190: 231-259.

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., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwich, J.R. & Ferrier, S. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. – Ecol. Appl. 19: 181-197.

Phillips, S.J., Dudík, M. & Schapire, R. 2004. A maximum entropy approach to species distribution modeling. – In: Anonymous (ed.), Proceedings of the 21st international conference on machine learning, ACM Press, New York, pp. 655-662.

Phillips, S. & Elith, J. 2010. POC-plots: calibrating species distribution models using presence-only data. – Ecology 91: 2476-2484.

*Pineda, E. & Lobo, J.M. 2009. Assessing the accuracy of species distribution models to predict amphibian species richness patterns. – J. anim. Ecol. 78: 182-190.

Platts, P.J., Ahrends, A., Gereau, R.E., McClean, C.J., Lovett, J.C., Marshall, A.R., Pellikka, P.K.E., Mulligan, M., Fanning, E. & Marchant, R. 2010. Can distribution models help refine inventory-based estimates of conservation priority? A case study in the Eastern Arc forests of Tanzania and Kenya. – Divers. Distrib. 16: 628-642.

*Raes, N. & ter Steege, H. 2007. A null-model for significance testing of presence-only species distribution models. – Ecography 30: 727-736.

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: 1689-1703.

*Rebelo, H. & Jones, G. 2010. Ground validation of presence-only modelling with rare species: a case study on barbastelles *Barbastella barbastellus* (Chiroptera: Vespertilionidae). – J. appl. Ecol. 47: 410-420.

Reineking, B. & Schröder, B. 2006. Constrain to perform: regularization of habitat models. – Ecol. Modelling 193: 675-690.

*Reside, A.E., Watson, I., VanDerWal, J. & Kutt, A.S. 2011. Incorporating low-resolution historic species location data decreases performance of distribution models. – Ecol. Modelling 222: 3444-3448.

*Riordan, E.C. & Rundel, P.W. 2009. Modelling the distribution of a threatened habitat: the California sage scrub. – J. Biogeogr. 36: 2176-2188.

Robertson, M.P., Cumming, G.S. & Erasmus, B.F.N. 2010. Getting the most out of atlas data. – Divers. Distrib. 16: 363-375.

*Rota, C.T., Fletcher, R., Jr., Evans, J.M. & Hutto, R.L. 2011. Does accounting for imperfect detection improve species distribution models? – Ecography 34: 659-670.

*Roura-Pascual, N., Brotons, L., Peterson, A.T. & Thuiller, W. 2009. Consensual predictions of potential distributional areas for invasive species: a case study of Argentine ants in the Iberian Peninsula. – Biol. Invasions 11: 1017-1031.

Roxburgh, S.H. & Mokany, K. 2010. On testing predictions of species relative abundance from maximum entropy optimisation. – Oikos 119: 583-590.

*Rupprecht, F., Oldeland, J. & Finckh, M. 2011. Modelling potential distribution of the threatened tree species Juniperus oxycedrus: how to evaluate the predictions of different modelling approaches? – J. Veg. Sci. 22: 647-659.

Rydgren, K., Halvorsen, R., Auestad, I. & Hamre, L.N. in press. Ecological design is more important than compensatory mitigation for successful restoration of alpine spoil heaps. – Rest. Ecol. in press.

Rydgren, K., Økland, R.H. & Økland, T. 2003. Species response curves along environmental gradients: a case study from SE Norwegian swamp forests. – J. Veg. Sci. 14: 869-880.

Santika, T. 2011. Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. – Global Ecol. Biogeogr. 20: 181-192.

Santika, T. & Hutchinson, M.F. 2009. The effect of species response form on species distribution model prediction and inference. – Ecol. Modelling 220: 2365-2379.

Schwarz, G. 1978. Estimating the dimension of a model. – Ann. Statist. 6: 461-464.

Segurado, P., Araújo, M.B. & Kunin, W.E. 2006. Consequences of spatial autocorrelation for niche-based models. – J. appl. Ecol. 43: 433-444.

*Sérgio, C., Figueira, R., Draper, D., Menezes, R. & Sousa, A.J. 2007. Modelling bryophyte distribution based on ecological information for extent of occurrence assessment. – Biol. Conserv. 135: 341-351.

Shipley, B. 2010. Community assembly, natural selection and maximum entropy models. – Oikos 119: 604-609.

Shipley, B., Vile, D. & Garnier, É. 2006. From plant traits to plant communities: a statistical mechanistic approach to biodiversity. – Science 314: 812-814.

Sokal, R.R. & Rohlf, F.J. 1995. Biometry, ed. 3. – Freeman, New York.

*Stachura-Skierczyńska, K., Tumiel, T. & Skierczyński, M. 2009. Habitat prediction model for three-toed woodpecker and its implications for the conservation of biologically valuable forests. – For. Ecol. Mgmt 258: 697-703.

Steyerberg, E.W., Eijkemans, M.J., Harrell Jr., F.E. & Habbema, J.D. 2000. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. – Statist. Med. 19: 1059-1079.

*Suárez-Seoane, S., García de la Morena, E.L., Prieto, M.B.M., Osborne, P.E. & de Juana, E. 2008. Maximum entropy niche-based modelling of seasonal changes in little bustard (*Tetrax tetrax*) distribution. – Ecol. Modelling 219: 17-29.

Suárez-Seoane, S., Osborne, P.E. & Rosema, A. 2004. Can climate data from METEOSAT improve wildlife distribution models? – Ecography 27: 629-636.

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

*Svenning, J.-C., Normand, S. & Kageyama, M. 2008. Glacial refugia of temperate trees in Europe: insights from species distribution modelling. – J. Ecol. 96: 1117-1127.

*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 Ecol. Biogeogr. 20: 904-914.

ter Braak, C.J.F. & Prentice, I.C. 1988. A theory of gradient analysis. – Adv. ecol. Res. 18: 271-317.

*Thompson, G.D., Robertson, M.-P., Webber, B.L., Richardson, D.M., Le Roux, J.J. & Wilson, J.R.U. 2011. Predicting the subspecific identity of invasive species using distribution models: *Acacia saligna* as an example. – Divers. Distrib. 17: 1001-1014.

*Thorn, J.S., Nijman, V., Smith, D. & Nekaris, K.A.I. 2009. Ecological niche modelling as a technique for assessing threats and setting conservation priorities for Asian slow lorises (Primates: Nycticebus). – Divers. Distrib. 15: 289-298.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. – J. R. Statist. Soc. Ser. B 58: 267-288.

Tingley, R. & Herman, T.B. 2009. Land-cover data improve bioclimatic models for anurans and turtles at a regional scale. – J. Biogeogr. 36: 1656-1672.

*Tinoco, B.A., Astudillo, P.X., Latta, S.C. & Graham, C.H. 2009. Distribution, ecology and conservation of an endangered Andean hummingbird: the violet-throated metaltail (*Metallura baroni*). – Bird Conserv. Int. 19: 63-76.

*Tittensor, D.P., Baco, A.R., Brewin, P.E., Clark, M.R., Consalvey, M., Hall-Spencer, J., Rowden, A.A., Schlacher, T., Stocks, K.I. & Rogers, A.D. 2009. Predicting global habitat suitability for stony corals on seamounts. – J. Biogeogr. 36: 1111-1128.

*Tognelli, M.F., Roig-Juñent, S.A., Marvaldi, A.E., Flores, G.E. & Lobo, J.M. 2009. An evaluation of methods for modelling distribution of Patagonian insects. – Revta chil. Hist. nat. 82: 347-360.

Trivedi, M.R., Berry, P.M., Morecroft, M.D. & Dawson, T.P. 2008. Spatial scale affects bioclimate model projections of climate change impacts on mountain plants. – Global Change Biol. 14: 1089-1103.

*Urbina-Cardona, J.N. & Flores-Villela, O. 2010. Ecological-niche modeling and prioritization of conservation-area networks for Mexican herpetofauna. – Conserv. Biol. 24: 1031-1041.

*Václávík, T. & Meentemeyer, R.K. 2009. Invasive species distribution modeling (iSDM): are absence data and dispersal constraints needed to predict actual distributions? – Ecol. Modelling 220: 3248-3258.

*VanDerWal, J., Shoo, L.P., Graham, C.H. & Williams, S.E. 2009a. Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know? – Ecol. Modelling 220: 589-594.

*VanDerWal, J., Shoo, L.P. & Williams, S.P. 2009b. New approaches to understanding late Quaternary climate fluctuations and refugial dynamics in Australian wet tropical rain forests. – J. Biogeogr. 36: 291-301.

van Neil, K.P. & Austin, M.P. 2007. Predictive vegetation modeling for conservation: impact of error propagation from digital elevation data. – Ecol. Appl. 17: 266-280.

Varela, S., Rodríguez, J. & Lobo, J.M. 2009. Is current climatic equilibrium a guarantee for the transferability of distribution model predictions? A case study of the spotted hyena. – J. Biogeogr. 36: 1645-1655.

Vaughan, I.P. & Ormerod, S.J. 2003. Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data. – Conserv. Biol. 17: 1601-1611.

*Veloz, S.D. 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. – J. Biogeogr. 36: 2290-2299.

Venables, W.N. & Ripley, B.D. 2002. Modern applied statistics with S. – Springer, New York.

*Verbruggen, H., Tyberghein, L., Pauly, K., Vlaeminck, C., van Nieuwenhuyze, K., Kooistra, W., Leliaert, F. & de Clerck, O. 2009. Macroecology meets macroevolution: evolutionary niche dynamics in the seaweed Halimeda. – Global Ecol. Biogeogr. 18: 393-405.

*Wang, Y., Xie, B., Wan, F., Xiao, Q. & Dai, L. 2007. The potential geographic distribution of Radopholus similis in China. – Agric. Sci. China 6: 1444-1449.

*Ward, D.F. 2007. Modelling the potential geographic distribution of invasive ant species in New Zealand. – Biol. Invasions 9: 723-735.

Ward, G., Hastie, T., Barry, S., Elith, J. & Leathwick, J.R. 2009. Presence-only data and the EM algorithm. – Biometrics 65: 554-563.

*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.

Warren, D.L. & Seifert, S.N. 2011. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. – Ecol. Appl. 21: 335-342.

*Webber, B.L., Yates, C.J., La Maitre, D.C., Scott, J.K., Kriticos, D.J., Ota, N., McNeill, A., Le Roux, J.J. & Midgley, G.F. 2011. Modelling horses for novel climate courses: insights from projecting potential distributions of native and alien Australian acacias with correlative and mechanistic models. – Divers. Distrib. 17: 978-1000.

*Weber, T.C. 2011. Maximum entropy modeling of mature hardwood forest distribution in four U.S. states. – For. Ecol. Mgmt 261: 779-788.

Whittaker, R.J., Araújo, M.B., Jepson, P., Ladle, R.J., J.E.M., W. & Willis, K.J. 2005. Conservation biogeography: assessment and prospect. – Divers. Distrib. 11: 3-23.

*Willems, E.P. & Hill, R.A. 2009. A critical assessment of two species distribution models: a case study of the vervet monkey (*Cercopithecus aethiops*). – J. Biogeogr. 36: 2300-2312.

*Williams, J.N., Seo, C.W., Thorne, J., Nelson, J.K., Erwin, S., O'Brien, J.M. & Schwartz, M.W. 2009. Using species distribution models to predict new occurrences for rare plants. – Divers. Distrib. 15: 565-576.

*Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A. & NCEAS Predicting Species Distributions Working Group 2008. Effects of sample size on the performance of species distribution models. – Divers. Distrib. 14: 763-773.

Wohlgemuth, T., Nobis, M.P., Kienast, F. & Plattner, M. 2008. Modelling vascular plant diversity at the landscape scale using systematic samples. – J. Biogeogr. 35: 1226-1240.

Wollan, A.K., Bakkestuen, V. & Halvorsen, R. 2011. Romlig prediksjonsmodellering av åpen grunnlendt kalkmark i Oslofjord-området. – Univ. Oslo NatHist. Mus. Rapp. 11: 176-196.

*Wollan, A.K., Bakkestuen, V., Kauserud, H., Gulden, G. & Halvorsen, R. 2008. Modelling and predicting fungal distribution patterns using herbarium data. – J. Biogeogr. 35: 2298-2310.

*Wolmarans, R., Robertson, M.P. & van Rensburg, B.J. 2010. Predicting invasive alien plant distributions: how geographical bias in occurrence records influences model performance. – J. Biogeogr. 37: 1797-1810.

Wood, S.N. 2006. Generalized additive models. – Chapman & Hall, London.

*Yates, C., McNeill, A., Elith, J. & Midgley, G. 2010. Assessing the impacts of climate change and land transformation on *Banksia* in the South West Australian floristic region. – Divers. Distrib. 16: 187-201.

*Yesson, C. & Culham, A. 2006. A phyloclimatic study of *Cyclamen*. – BMC evol. Biol. 6: 72: 1-23.

*Yost, A.C., Petersen, S.L., Gregg, M. & Miller, R. 2008. Predictive modeling and mapping sage grouse (*Centrocercus urophasianus*) nesting habitat using maximum entropy and a long-term dataset from southern Oregon. – Ecol. Informatics 3: 375-386.

*Young, B.F., Franke, I., Hernandez, P.A., Herzog, S.K., Paniagua, L., Tovar, C. & Valqui, T. 2009. Using spatial models to predict areas of endemism and gaps in the protection of Andean slope birds. – Auk 126: 554-565.

Zielinski, W.J., Dunk, J.R., Yaeger, J.S. & LaPlante, D.W. 2010. Developing and testing a landscape-scale habitat suitability model for fisher (*Martes pennanti*) in forests of interior northern California. – For. Ecol. Mgmt 260: 1579-1591.

Zuur, A.F., Ieno, E.N. & Smith, G.M. 2007. Analysing ecological data. – Springer, New York.

APPENDIX 1: TERMINOLOGY

Notation, terms and explanation of terms used in this paper and comparison with other papers. Bold-face italicised capital letters are used to denote vectors with N elements, i.e., that are defined for all grid cells in the study area, and bold-face normal letters are used for matrices. P = Phillips et al. (2006) and Phillips & Dudík (2008); E = Elith et al. (2011).

Terminology used in this paper			Terminology in previous papers on MaxEnt distribution modelling		
Notation	Term	Explanation	Notation P	Notation E	Term, if different from term used in this paper
a	# of factor levels	number of factor levels into which a categorical explanatory variable is divided			
$\boldsymbol{B}$	observed presence or absence (OPA) vector	presence or absence in the N grid cells in $\boldsymbol{D}$; $\boldsymbol{B} = [b_1, ..., b_i, ..., b_N]^T$ is a vector representation of binary true presence/absence (P/A) data	y	y	P, E: use y to denote the vector of true (but unknown) presence or absence
$\overline{b}$	prevalence	the mean of the observed presence or absence (OPA) vector $\boldsymbol{B}$ over the N grid cells			
b_i	presence/absence value	characteristic of grid cell i in $\boldsymbol{D}$ as given by the observed presence or absence (OPA) vector $\boldsymbol{B}$: presence ($b_i = 1$) or absence ($b_i = 0$)			
$\boldsymbol{C}$	observed presence (OP) vector	observed presence (or lack of information on presence or absence) in the N grid cells in $\boldsymbol{D}$; $\boldsymbol{C} = [c_1, ..., c_i, ..., c_N]$T is a vector representation of discrete presence-only (PO) data			
$\overline{c}$	frequency of presence	the mean of the observed presence (OP) vector $\boldsymbol{C}$ over the N grid cells			
c_i	presence or lack of information on presence or absence	characteristic of grid cell i in $\boldsymbol{D}$ as given by the presence-only vector $\boldsymbol{C}$: observed presence ($c_i = 1$) or uninformed background ($c_i = 0$)			
$\boldsymbol{D}$	set of grid cells used for MaxEnt modelling	the set of N grid cells into which the study area used is rasterised; observation units in a distribution modelling study; $\boldsymbol{D} = \{d_1, ..., d_i, ..., d_N\}$	X	L	E: the landscape of interest

Appendix 1 (Continued).

Terminology used in this paper			Terminology in previous papers on MaxEnt distribution modelling		
Notation	Term	Explanation	Notation P	Notation E	Term, if different from term used in this paper
$\boldsymbol{D}_+$	subset of observed presence, or presence, grid cells	the subset of n grid cells in $\boldsymbol{D}$ in which the modelled target is observed present, i.e., for which $c_i = 1$, or the subset of n_+ cells for which the target is present, i.e., for which $b_i = 1$		L_1	
$\boldsymbol{D}_-$	subset of observed absence, or uninformed background, grid cells	the subset of n grid cells in $\boldsymbol{D}$ in which the modelled target is observed absent, i.e., for which $c_i = 0$, or the subset of n_- cells for which the target is absent, i.e., for which $b_i = 0$			
D_e	set of grid cells used for evaluation of a distribution model	the set of grid N_e cells used for evaluating a distribution model; $D_e = \{d_{e1}, ..., d_{ei}, ..., d_{eN_e}\}$			
$\boldsymbol{D}_l$	subset of grid cells with specific values for the explanatory variables	the subset of grid cells in $\boldsymbol{D}$ for which the s explanatory variables take on values $Z_l = [z_{l1}, ..., z_{lj}, ..., z_{ls}]^T$; where each z_{lj} can be an exact value or an interval	$X(z)$		
$\boldsymbol{D}_T$	set of all grid cells in the study area	the set of all N_T grid cells in the study area; $\boldsymbol{D}_T = \{d_1, ..., d_i, ..., d_{N_T}\}$			
d_i	grid cell	the ith among N grid cells in the set $\boldsymbol{D}$ of grid cells in the study area	x	x	P: point E: pixel, location
f	unconditional probability density for a derived variable vector	the unconditioned probability density for a vector X_l of values for all m derived variables recorded at site l		$f(\mathbf{z})$	E: the probability density of ‘covariates’ across L
f_0	probability density for the derived variable vector, conditioned on absence	probability density for the vector X_l of values for all m derived variables at site l, conditioned on absence of the modelled target ($b_l = 0$)		$f_0(\mathbf{z})$	E: the probability density of ‘covariates’ across locations within L where the modelled target is absent
f_1	probability density for the derived variable vector, conditioned on presence	probability density for the vector X_l of values for all m derived variables at site l, conditioned on presence of the modelled target ($b_l = 1$)		$f_1(\mathbf{z})$	E: the probability density of ‘covariates’ across locations within L where the modelled target is present

Appendix 1 (Continued).

Terminology used in this paper			Terminology in previous papers on MaxEnt distribution modelling		
Notation	Term	Explanation	Notation P	Notation E	Term, if different from term used in this paper
g	MaxEnt function of derived variables	the function by which MaxEnt raw output q_l, is estimated for a vector X_l of values for all m derived variables (in site l); g_Θ (X_l)			
g'	MaxEnt function of (raw) explanatory variables	the function by which MaxEnt raw output q_l, is estimated for a vector Z_l of values for all s explanatory variables (in site l); g_Θ' (Z_l)			
h_k; h_K	transformation function	the function by which one or more of the explanatory variables in $\mathbf{Z}$ are transformed into derived variables $\boldsymbol{X}_k$ or, more generally, a derived variable of type K			
h_j^{-1}	back-transformation function	the function by which one or more of the derived variables in $\mathbf{X}$ are back-transformed into explanatory variable Z_j			
i	grid cell index	index for grid cells d, organised with $i = 1, \dots , n$ being the n presence cells and $i = n + 1, \dots , N$ the N – n uninformed background cells			
i_0	one specific grid cell	one specific presence grid cell d with given properties, e.g., one grid cell selected at random from all presence cells			
j	explanatory variable index	index for explanatory variables, running from 1 to s			
K	index for type of derived variable	index for type of derived variable; of which nine are recognised; C = binary set derived derived from a categorical EV; D = deviation; H = hinge, L = linear; M = monotonous; O = covariance; P = product; T = threshold; X = complex spline			
k	derived variable index	index for derived variables, running from 1 to m	j		
L_t; L_Θ	likelihood for MaxEnt model t or the MaxEnt model given by Θ	the likelihood for MaxEnt models denoted by Q_t or Q_Θ (i.e., the model with parameter vector Θ)			
L_{t+}; $L_{\Theta+}$	likelihood for presence cells for MaxEnt model t or the MaxEnt model given by Θ	the likelihood for the n presence cells for MaxEnt models denoted by Q_t or Q_Θ (i.e., the model with parameter vector Θ)			

Appendix 1 (Continued).

Terminology used in this paper			Terminology in previous papers on MaxEnt distribution modelling		
Notation	Term	Explanation	Notation P	Notation E	Term, if different from term used in this paper
L_{t-}; $L_{\Theta-}$	likelihood for uninformed background cells for MaxEnt model *t* or the MaxEnt model given by Θ	the likelihood for the *N*–*n* uninformed background cells for MaxEnt models denoted by Q_t or Q_Θ (i.e., the model with parameter vector Θ)			
ln L_{t+}; ln $L_{\Theta+}$	log likelihood for presence cells	the sum of natural logarithms for the *n* presence cells for MaxEnt models denoted by Q_t or Q_Θ (i.e., the model with parameter vector Θ)			
ln L_t; ln L_Θ	log loss	the negative log likelihood for the *n* presence cells divided by *n*, for MaxEnt models denoted by Q_t or Q_Θ (i.e., the model with parameter vector Θ)			
ln Λ_t; ln Λ_Θ	penalised log loss, regularised log loss	log loss for MaxEnt models denoted by Q_t or Q_Θ (i.e., the model with parameter vector Θ), penalised by the magnitude of the parameters			
l	site index	index for sites of potential relevance to a distribution modelling project; the site being defined by its vector of values for explanatory variables Z_l (or the derived vector of values for derived variables X_l); sites of potential relevance comprise the *N* grid cells in ***D*** and all combinations of explanatory variable values within the range of observed values for each of the s explanatory variables			
m	# of derived variables	the total number of derived variables used for MaxEnt modelling; total number of parameters in a MaxEnt model minus 1	*n*		P, E: the term 'feature' is used for 'derived variable'
m_j	# of factor levels recorded for a categorical explanatory variable	the number of factor levels u recorded for a categorical explanatory variable Z_j; corresponds to the number of binary variables derived from Z_j			
N	# of grid cells used for MaxEnt modelling	the number of grid cells used as observations (presence + uninformed background cells) in a distribution modelling study (i.e., the number of elements in D)	*N*		

Appendix 1 (Continued).

Terminology used in this paper			Terminology in previous papers on MaxEnt distribution modelling		
Notation	Term	Explanation	Notation P	Notation E	Term, if different from term used in this paper
N_e	# of grid cells in (independently collected) evaluation data set	the total number of grid cells in a data set of presence/absence observations, collected independently of the data used to parameterise the model, the intended use of which is to evaluate a distribution model			
N_T	# of grid cells in the study area	the total number of grid cells in the study area (i.e., the number of elements in $\boldsymbol{D}_T$)			
n	# of observed presence cells	the total number grid cells in ***D***, for which presence is recorded according to the PO vector ***C***	m	m	
n+	# of presence cells	the total number grid cells in ***D***, for which presence is recorded according to the P/A vector ***B***			
$N - n$	# of uninformed background cells	total number of grid cells in ***D***, for which presence is not recorded according to the P/A vector ***B***			
$N - n$+	# of absence cells	the total number grid cells in ***D***, for which absence is recorded according to the P/A vector ***B***			
n_a	# of correctly predicted presences	the number of presence cells in an evaluation data set correctly predicted by use of binary model output $\tilde{Q}_{q_0}$			
n_b	# of incorrectly predicted presences	the number of absence cells in an evaluation data set for which presence is incorrectly predicted by use of binary model output $\tilde{Q}_{q_0}$			
n_c	# of incorrectly predicted absences	the number of presence cells in an evaluation data set for which absence is incorrectly predicted by use of binary model output $\tilde{Q}_{q_0}$			
n_d	# of correctly predicted absences	the number of absence cells in an evaluation data set correctly predicted by use of binary model output $\tilde{Q}_{q_0}$			
n_f	# of grid cells with vector of values for derived variable X_f	the total number of grid cells in ***D*** with vector of values for derived variable exactly equal to X_f			

Appendix 1 (Continued).

Terminology used in this paper			Terminology in previous papers on MaxEnt distribution modelling		
Notation	Term	Explanation	Notation P	Notation E	Term, if different from term used in this paper
Q	MaxEnt model	a MaxEnt model			
$\boldsymbol{Q}$	vector of MaxEnt model estimates; the MaxEnt distribution	MaxEnt 'raw' estimates (predictions) of the probability for π_i for all grid cells i in $\boldsymbol{D}$, $\boldsymbol{Q} = [q_1, ..., q_i, ..., q_N]^{\mathrm{T}}$; $\boldsymbol{Q}$ is the discrete response probability distribution modelled by MaxEnt			
Q^*	vector of MaxEnt model estimates for observed presence observations	MaxEnt 'raw' estimates (predictions) of the probability for π_i; $Q^* = [q_1, ..., q_i, ..., q_n]^{\mathrm{T}}$, for the n observed presence cells ($i \leq n$)			
Q_t; Q_{t-}	vector of MaxEnt model estimates for two models	'raw' estimates (predictions) of the probability for π_i for all grid cells i in $\boldsymbol{D}$, for two MaxEnt models, typically used for two models of which one, Q_{t-}, is a submodel of the other, Q_t			
Q_S	saturated MaxEnt model	MaxEnt reference model that perfectly predicts observed presences ($c_i = 1$) and predicts absence for uninformed background cells ($c_i = 0$)			
Q_0	MaxEnt null model	MaxEnt reference model that explains no variation in the dependent variable, typically containing no derived variables			
Q_θ	the MaxEnt model given by $\boldsymbol{\Theta}$	MaxEnt model with parameters given by the model parameter vector $\boldsymbol{\Theta}$			
$\dot{Q}$	MaxEnt model, probability ratio output	MaxEnt model Q estimates for grid cell i or site t ($\dot{q}_i$ or $\dot{q}_t$), interpretable as unbiased estimates of the ratio f_1/f			
$\ddot{Q}$	MaxEnt model, cumulative output	MaxEnt model Q estimates for grid cell i, interpretable as the cumulative probability value of the MaxEnt distribution $\boldsymbol{Q}$ in grid cell i			
$\dddot{Q}$	MaxEnt model, logistic output	MaxEnt model Q estimates, scaled to range [0,1] and with value for $X^* = [\overline{x}_1^*, ..., \overline{x}_k^*, ..., \overline{x}_m^*]$ or, alternatively, $Z^* = [z_1^*, ..., z_j^*, ..., z_s^*]^{\mathrm{T}}$, equal to a parameter τ			

Appendix 1 (Continued).

Terminology used in this paper			Terminology in previous papers on MaxEnt distribution modelling		
Notation	Term	Explanation	Notation P	Notation E	Term, if different from term used in this paper
$\check{Q}$	MaxEnt model, probability-of-presence output	MaxEnt model estimates, scaled to range [0,1] and calibrated by use of independent calibration and evaluation data to be interpretable as unbiased estimates of the true probability of presence in a site of unit size			
$\tilde{Q}_{q_0}$	binary predictions from a MaxEnt model, using q_0 as prediction threshold	MaxEnt output, reclassified by use of q_0 as prediction threshold, to a binary variable with outcomes: predicted presence ($\tilde{q}_i = 1$) and predicted absence ($\tilde{q}_i = 0$)			
q_i; q_l; $\hat{\pi}_i$	MaxEnt model estimate	MaxEnt ‘raw’ estimates, for a site $i \in \boldsymbol{D}$, $q_i = \Pr(i = i_0 \mid b_{i_0} = 1; \Theta, \mathbf{Z}) = \hat{\pi}_i$, for an arbitrary site l interpretable as the relative predicted probability of presence (RPPP) as modelled by MaxEnt, $q_l = g_\theta(X_l)$	$q_\lambda(x)$; $\hat{\pi}(\mathrm{x})$		
$\dot{q}_i$; $\dot{q}_l$	MaxEnt probability ratio output value	MaxEnt model Q estimate given in probability-ratio output format (cf. $\dot{Q}$) for grid cell i in $\boldsymbol{D}$ or site l			
$\ddot{q}_i$; $\ddot{q}_l$	MaxEnt cumulative output value	MaxEnt model Q estimate given in cumulative output format (cf. $\ddot{Q}$) for grid cell i in $\boldsymbol{D}$ or site l			
$\dddot{q}_i$; $\dddot{q}_l$	MaxEnt logistic output value	MaxEnt model Q estimate given in logistic output format (cf. $\dddot{Q}$) for grid cell i in $\boldsymbol{D}$ or site l			
$\check{q}_i$; $\check{q}_l$	MaxEnt probabilty-of-presence output value	MaxEnt model Q estimate given in probability-of-presence output format (cf. $\check{Q}$) for grid cell i in $\boldsymbol{D}$ or site l			
$\tilde{q}_i$; $\tilde{q}_l$	MaxEnt binary prediction output value	MaxEnt model estimate, reclassified to a binary prediction of presence ($\tilde{q}_i = 1$) or absence ($\tilde{q}_i = 0$)			
q_0	prediction threshold	value used to split MaxEnt model estimates q_i in groups of predicted presences and predicted absences (cf. $\tilde{Q}_{q_0}$)			
s	# of explanatory variables	the total number of explanatory variables used in a DM study			

Appendix 1 (Continued).

Terminology used in this paper			Terminology in previous papers on MaxEnt distribution modelling		
Notation	Term	Explanation	Notation P	Notation E	Term, if different from term used in this paper
t	model index	index used to denote an arbitrary MaxEnt model			
U	# of randomisations	the total number of randomisations (permutations) of an explanatory variable or an derived variable used in a randomisation test			
U_0	# of permutations for which a test statistic exceeds (or is lower than, depending on the context) the value obtained for the data	the total of permutations for which a test statistic exceeds (or is lower than; applies to cases where lower values of the statistic indicates better performance) the value obtained for the data			
u	index used for several purposes	index for study areas, permutations, etc.			
v_t	variation accounted for	the variation accounted for by MaxEnt model t, given in log loss units			
$\acute{v}_t$	regularised variation accounted for	the variation accounted for by MaxEnt model t with ℓ_1-regularisation, given in log loss units			
V_t	fraction of total variation accounted for	the variation accounted for by MaxEnt model t expressed as the fraction of the total explainable variation			
$\acute{V}_t$	regularised fraction of total variation accounted for	the variation accounted for by MaxEnt model t with ℓ1-regularisation, given in log loss units, expressed as the fraction of the total explainable variation			
w_t	residual variation	residual variation not accounted for by MaxEnt model t			
$\mathbf{X}$	derived variables matrix	$N \times m$ matrix with values for m derived variables, $\boldsymbol{X}_1, ..., \boldsymbol{X}_k, ..., \boldsymbol{X}_m$, for each of the N grid cells in $\boldsymbol{D}$ (the derived variables are derived from s explanatory variables by transformation)			
X_f	vector of values for derived variables	vector of values for the m derived variables in grid cells in $\boldsymbol{D}$; X_f accurately characterises a set of n_d grid cells with exactly this specific combination of values for all derived variables			

Appendix 1 (Continued).

Terminology used in this paper			Terminology in previous papers on MaxEnt distribution modelling		
Notation	Term	Explanation	Notation P	Notation E	Term, if different from term used in this paper
X_i; X_l	vector of values for derived variables	values for the m derived variables in grid cell *i*; $X_i = [x_{i1}, ..., x_{ik}, ..., x_{im}]$, row vector in **X**; or in site *l*, X_l			
$\boldsymbol{X}_k$	derived variable vector	values for derived variable *k* in the *N* grid cells in ***D***; $\boldsymbol{X}_k = [x_{1k}, ..., x_{ik}, ..., x_{Nk}]^T$ is a column vector in **X**	f; g; f_j		P, E: the term 'feature' is used for 'derived variable'
$\boldsymbol{X}_k^*$	derived variable vector for observed presence observations	values for derived variable *k* in the *n* observed presence cells *i* ($i \leq n$) in ***D***+; $\boldsymbol{X}_k^* = [x_{1k}, ..., x_{ik}, ..., x_{nk}]^T$			
X^*	vector of mean values for all derived variables in observed presence cells	$X^* = [\bar{x}_1^*, ..., \bar{x}_k^*, ..., \bar{x}_m^*]$		$h(\mathbf{z})$	E: vector of 'features'
x_{ki}; x_{kl}	derived variable value	the value for derived variable *k* in grid cell *i* in ***D***, x_{ik}, or in site *l*, x_{lk}	f_j (x)		
$\bar{x}_k$	overall mean of derived variable *k*	the mean of derived variables vector $\boldsymbol{X}_k$; i.e., the mean of x_{ik} over all *N* grid cells in ***D***	|X|		
$\bar{x}_k^*$	mean of derived variables *k* in observed presence cells	the mean of x_{ik} over the *n* observed presence grid cells in ***D***	$\tilde{\pi}[f_j]$		P: the empirical average of 'feature' f_j
$\tilde{x}_k^*$	weighted average of derived variables *k* using model output as weights	the average of derived variables *k* over all *N* grid cells in ***D*** using raw output $\boldsymbol{Q} = [q_1, ..., q_i, ..., q_N]^T$ from a MaxEnt model as weights	$\hat{\pi}[f_j]$		P: the expectation of 'feature' fj under $\hat{\pi}$; $\pi[f_j]$ is used to denote the expectation of 'feature' f_j under π
$\boldsymbol{Y}$	binary response variable used in distribution modelling	a general binary variable, with values for the *N* grid cells in ***D***, used as response variable in distribution modelling; $\boldsymbol{Y} = [y_1, ..., y_i, ..., y_N]^T$; the y_i can be of the presence-only or presence/absence type of data			
y_i	response value	the value for the response variable ***Y*** in cell *i* in ***D***			
$\mathbf{Z}$	explanatory variables matrix	$N \times s$ matrix with observations of *s* explanatory variables, $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_j, ..., \boldsymbol{Z}_s$, in each of the *N* grid cells in ***D***			

Appendix 1 (Continued).

Terminology used in this paper			Terminology in previous papers on MaxEnt distribution modelling		
Notation	Term	Explanation	Notation P	Notation E	Term, if different from term used in this paper
Z_i; Z_l	vector of values for explanatory variables	observed values for the *s* explanatory variables in cell *i*; $Z_i = [z_{i1}, ..., z_{ij}, ..., z_{is}]$ is a row vector in **Z**; or in site *l*, Z_l			
Z_j	explanatory variable vector	observed values for explanatory variable *j* in the *N* grid cells in ***D***; $Z_j = [z_{j1}, ..., z_{ji}, ..., z_{jN}]^{\mathrm{T}}$ is a column vector in **Z**		z	E: vector of environmental 'covariates'
Z^*	vector of mean values for all explanatory variables in observed presence cells	$Z^* = [\overline{z}_1^{\,*}, ..., \overline{z}_j^{\,*}, ..., \overline{z}_s^{\,*}]$			
$\boldsymbol{z_{ji}}$	explanatory variable value	the observed value of explanatory variable *j* in grid cell *i* in ***D***	v_i		P: values taken on by a categorical explanatory variable
$\overline{z}_j$	overall mean of explanatory variable *j*	the mean of explanatory variable vector Z_j; i.e., the mean of z_{ij} over all *N* grid cells in ***D***			
η	# of (independent) observations; parameter used for F-ratio test	the number of (independent) observations; used for determination of the degrees of freedom in model-comparison *F*-ratio test			
$\boldsymbol{\Theta}$	model parameter vector	parameters of a MaxEnt distribution model *Q*; $\Theta = [\theta_0, \theta_1, ..., \theta_k, ..., \theta_m]$	λ	β	P: 'feature weights', 'model coefficients' E: vector of coefficients Note that terms of P & E do not include θ_0)
$\boldsymbol{\Theta}_t$	parameter vector for model *t*	parameters of one specific distribution model Q_t			
θ_k	model parameter	parameter *k* in distribution model ($k = 0, 1, ..., m$)	λ_j		P: 'feature weight', 'model coefficient'
θ_0	model intercept	intercept, a parameter of the distribution model *Q* that does not depend on the derived variables **Z**	$-\ln Z_\lambda$	α	P: Z_λ is referred to as 'normalizing constant' E: α is referred to as 'normalizing constant'

Appendix 1 (Continued).

Terminology used in this paper			Terminology in previous papers on MaxEnt distribution modelling		
Notation	Term	Explanation	Notation P	Notation E	Term, if different from term used in this paper
λ	regularisation parameter	a constant set by the user to be used in ℓ1-regularisation	β	λ	
λ_K	regularisation parameter specific to a type of derived variables	a constant set by the user to be used in ℓ1-regularisation, specific for derived variables of type K	β_j	λ_j	
$\boldsymbol{\Pi}$	probability vector for selecting one specific presence grid cell given that the identity of presence cells is known	the probability that a randomly selected presence grid cell $i0$: $c_{i0} = 1$, is cell i; $\boldsymbol{\Pi} = [\pi_1, ..., \pi_i, ..., \pi_N]^T$ where $\pi_i = \frac{1}{N}$ and $\pi_1 = 0$ otherwise	π		P: 'the unknown probability distribution we want to approximate'
$\boldsymbol{\Pi}_0$	probability vector for selecting one specific presence grid cell given that the identity of presence cells is not known	the probability that a randomly selected presence grid cell i_0: $c_{i0} = 1$, is cell i, given that the identity of presence cells is not known; $\boldsymbol{\Pi} = [\pi_1, ..., \pi_i, ..., \pi_N]^T$ where $\pi_i = \frac{1}{n}$	π		P: 'the unknown probability distribution we want to approximate'
π_i	the probability that one specific presence grid cell is a given cell	the probability that a randomly selected presence cell i_0: $c_{i0} = 1$, is cell i	$\pi(x)$		
$\hat{\pi}_i$	model estimate for the probability that one specific presence cell is a given cell	the probability that a randomly selected presence cell i_0: $c_{i0} = 1$, is cell i, as estimated by a distribution model; $\hat{\pi}_i = q_i$ for all grid cells in $\boldsymbol{D}$	$\hat{\pi}(\mathrm{x})$		
τ	logistic output parameter	a parameter chosen by the user to be the fixed value of the MaxEnt logistic output value for an 'average presence site' $X^* = (x_1^*, ..., x_k^*, ..., x_m^*)$ or, alternatively, $Z^* = (z_1^*, ..., z_j^*, ..., z_s^*)$		τ	
			h		P: knot, relevant for derived variables of types H and T
			k	k	P: the number of 'binary features' created from a categorical explanatory variable

Appendix 1 (Continued).

Terminology used in this paper			Terminology in previous papers on MaxEnt distribution modelling		
Notation	Term	Explanation	Notation P	Notation E	Term, if different from term used in this paper
			$H(\hat{\pi})$		P: the entropy of $\hat{\pi}$
			$s(\mathbf{z})$		E: sample selection bias
			$s^2[f_j]$		P: empirical variance of 'feature' f_j
			Z_λ		P normalising constant; corresponds to $e^{-\theta}$
			$\eta(\mathbf{z})$		E: defined as ln $(f_1(\mathbf{z})/f(\mathbf{z}))$ or, alternatively, as $\alpha + \beta \bullet h(\mathbf{z})$

APPENDIX II: INDEX

For bold face italicised terms, explicit definitions are given in the text [and, eventually, also in Appendix I of Halvorsen (2012)], bold-face page number refer to page on which the definition is given; plain bold-face letters refer to terms only defined in Appendix I of Halvorsen (2012).