MethodsX 8 (2021) 101394

Contents lists available at ScienceDirect

MethodsX

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

Method Article

SDM-CropProj – A model-assisted framework to forecast crop environmental suitability and fruit production

Salvador Arenas-Castro[a,b,c,*], João Gonçalves[a,c,d]

[a] Escola Superior Agrária, Instituto Politécnico de Viana do Castelo, Praça Gen. Barbosa 44, 4900-347 Viana do Castelo, Portugal
[b] Área de Ecología, Dpto. de Botánica, Ecología y Fisiología Vegetal, Facultad de Ciencias, Universidad de Córdoba, Campus de Rabanales, 14071 Córdoba, España
[c] InBIO/CIBIO - Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, Campus Agrário de Vairão, Rua Padre Armando Quintas, 4485-601 Vairão, Portugal
[d] proMetheus - Instituto Politécnico de Viana do Castelo, Rua Escola Industrial e Comercial Nun'Álvares, 4900-347 Viana do Castelo, Portugal

ABSTRACT

The effects of climate change (CC) will impact species ranges, and crops are no exception. Anticipating these changes through forecasting the environmentally suitable area of crops would help to reduce or mitigate the impact and adapt ecological and economic strategies. To forecast the CC effects on crops, we describe here a model-assisted framework (hereafter SDM-CropProj) that combines two modelling steps to be implemented in sequence: i) a multi-technique calibration process and ensemble-forecasting approach to predict the current and future environmental suitability of target crops; ii) a parsimonious univariate log-log linear model to relate the average total annual production to the current SDM-based suitable area. Different metrics for assessing the model's predictive performance showed that:

- Crop production is related to model-predicted suitable area, thus allowing to obtain future projections of total fruit production based on climate scenarios.
- The SDM-CropProj framework can assess potential pathways and trends in annual production due to changes in the environmental suitability and the distribution of multiple crop varieties/types as a consequence of CC, offering insights to other areas and crop types.

DOI of original article: 10.1016/j.scitotenv.2019.136161

* Corresponding author.
E-mail address: b62arcas@uco.es (S. Arenas-Castro).

https://doi.org/10.1016/j.mex.2021.101394

ARTICLE INFO
Method name: Environmentally-suitable predicted area to forecast fruit crop production
Keywords: Agroecosystems, Climate change, Crop production, Long-term projections, Machine learning, Species distribution models
Article history: Received 28 April 2021; Accepted 21 May 2021; Available online 23 May 2021

Specifications table

Subject Area:	Bioinformatics
More specific subject area:	Climate change projections of fruit crop production
Method name:	Environmentally-suitable predicted area to forecast fruit crop production
Name and reference of original method:	Not applicable
Resource availability:	https://github.com/salvador-arenas-castro/SDM-CropProj.git

Introduction

Although Species Distribution Models (SDMs) are commonly used for modeling animal and plant distributions in natural environments [1], SDMs are also increasingly used to model crops [2–4]. This modelling approach has become a reliable tool to predict the current and potential distribution of target crops, as well as to obtain projections of future environmental suitability in the face of climate change (CC). At the same time, SDMs coupled with production data can go even further and be employed to obtain projections of crop production according to different CC scenarios.

Method details

The spatial and temporal SDM-CropProj framework combines two modelling steps: i) a multi-technique calibration process and ensemble-forecasting approach based on SDMs; ii) a univariate log-log linear model to relate crop production to SDM-based suitable area. This cross-scale approach is expected to be widely applicable to different crop types worldwide, as well as a broad range of agroecosystems. Thus, the input data must be selected as convenient according to objectives but complying with guidelines described in the following subsections. Please find detailed instructions and a ready-to-use example at: https://github.com/salvador-arenas-castro/SDM-CropProj.git based on data for olive crop production in Andalusia (Spain).

Input data

The SDM-CropProj framework requires geographic location data for cultivated varieties (or any crop type), including presence records and, if available, absences too – as response variable – along with environmental variables as raster layers covering the whole region of interest – as predictive variables. These data must have similar spatial accuracy and resolution and the same coordinate system.

Presence/absences records for each target crop

Presences refer to spatial points (or raster cells) with at least one observation of the target crop variety/type (i.e., suitable locations), while absences relate to points with no observations (i.e., unsuitable areas) [5]. If true absences are not available, pseudo-absences (absence points created artificially) can also be used as surrogates (see below). Crop records must be resampled to match the spatial resolution of available environmental variables (ca. 1 km pixel size; see example below). To do so, both R and QGIS open-source software offer various tools for that purpose (e.g., resampling to cell centroid). This procedure also enables the removal of duplicate records for the same raster cell. Overall, it must be considered that the input data, in this case, the occurrence data, may harbor potential geographical or nomenclature biases (e.g., duplicates, errors in coordinates, misidentified species/varieties, etc.) that can influence the results of models. For that reason, data tidying and filtering procedures must be used before model training stages [6]. R packages such as 'dplyr' or 'tidyr' offer functions to handle these issues.

Predictive variables

Environmental variables (i.e., predictive variables) are used to explain the environmental suitability and distribution of the response variable (presences in this case). These should be selected based on previous (auto-)ecological knowledge of the species/varieties (from literature review or expert-based). Variable selection usually entails choosing a core set of non, or minimally correlated factors that control or limit the distribution or environmental suitability of the target crop species/variety at several scales, from coarse (e.g., climate) to local factors (e.g., topography, soil attributes). SDM accuracy generally improves with increasing numbers of variables until an asymptote is reached [7]. However, the number of variables should be much less than the number of records (the Harrelĺs rule: one variable per ten or twenty presence records has been suggested) [8]; thus, the fewer species records are available, the fewer variables should be introduced in the model [9,10]. In our example, variables related to plant physiology and distribution, such as water balance, (bio-)climatic variables, and soil properties, were selected as predictors of crop suitability and distribution. To match the spatial resolution of raster variables to presence data, all variables were re-projected to the same geographic coordinate system (WGS 1984) and resampled from their original spatial resolutions to 1×1 km pixel size by using the function 'aggregate' (available in the 'raster' package in R v.4.0). 'Aggregate' is a GIS function that groups rectangular areas to create larger pixels by applying a mathematical function (e.g., mean, maximum, minimum, sum, range). In our specific case the average function was employed for spatial aggregation.

Most modelling algorithms implemented in the SDM-CropProj framework are sensitive to high levels of correlation among predictor variables [11]. A multicollinearity reduction procedure is often applied to simplify models to improve model parsimony, reduce overfitting, and decrease overall correlation between predictors (leading to less parameter estimation uncertainty). For instance, Variance Inflation Factors (VIF) produce a diagnosis of collinearity among variables useful for reducing the variables' set. For initial screening, and considering that the correlation matrix is a good indicator of multicollinearity, which signals the need for further investigation, we recommend calculating the non-parametric Spearman correlation (r) matrix by using the function 'cor' in 'stat' v4.0 R package and remove predictors with $|r|>0.8$ (i.e., highly correlated variables). Typically, those predictors that better represent extreme ecological limiting factors among all varieties are selected. Often in SDM development, bioclimatic variables calculated from long-term time series (i.e., 30-years long) are used to portray extreme ecological limiting factors (e.g., minimum temperature of the coldest month, maximum temperature of the warmest month). However, it is also recommended to incorporate annualized climate variables or climatic extremes if these data are available because they impact and report on the phenology of plants in a more explicit way. Finally, to perform the future projections, there are distinct useful portals that provide datasets for past and future climate scenarios (e.g., WorldClim: https://worldclim.org; CHELSA: https://chelsa-climate.org). Nevertheless, there are associated limitations concerning the quality of these data sets due to the relatively coarse and uneven density of the underlying weather stations network from which the gridded data sets are derived. In our example study, different regional climate change scenarios specifically developed for the study region at 200×200 m grid cells [12] were selected to understand the future distribution and assess the environmental suitability of each crop variety. The regional scenarios were produced based on the Third Generation Coupled Global Climate Model (MCGs; CNCM3) for a balance across all sources (A1b; IV IPCC Report) and for three periods: 2040, 2070, and 2100. Current (bio-)climatic variables were calculated as average values for the reference period of 1961-2000.

Crop(s) production data

One of the main outcomes from the SDM-CropProj framework is the forecast of crop production. For this purpose, regionalized crop production data (i.e., by municipality, province, state) for one specific year or multiple years are needed. However, as different environmental factors can produce inter-annual fluctuations, it is recommended to average multiple years of production values to minimize these effects. Hence, this averaged total annual production per region is considered the response variable in crop productivity modelling. A raster layer with region integer codes (one for each unit) is also required (check example for details).

Modelling procedure

The SDM-CropProj framework combines two modelling steps to be implemented in sequence.

SDM calibration, evaluation, and ensemble projections

The SDM-CropProj method is based on an ensemble-forecasting approach, implemented through the 'biomod2' R package [13]. Biomod2 is a complete suite of tools for running, evaluating, and interpreting SDMs, including several presence-absence and presence-only algorithms and several model evaluation metrics. One of the strongest points of this package is the advanced ways to ensemble all models: by algorithm, by pseudo-absence datasets, by repetitions, or by combining algorithms, datasets, or repetitions. The SDM-CropProj framework is fitted by all ten modelling techniques available for each set of models and using hyper-parameters set by default (but changeable according to user needs). Since true-absence data were not available for the target crop varieties, pseudo-absences are often generated. In our example, ten different sets of randomly distributed unoccupied grid cells for the study region are generated, with the following constraints: 1) a prevalence of 0.5 for each variety, which means that presences and absences will be equally weighted to avoid potential bias caused by different levels of prevalence in the presence/absence datasets; and 2) defining a minimum distance between pseudo-absences, corresponding with each grain size (1 km), and without overlapping with presences, to avoid spatial autocorrelation and in order to cover the different ecological conditions in each study area. Model evaluation scores are calculated by employing hold-out cross-validation and running the whole process ten times for each set of pseudo-absences (this value can be changed if needed). Species/varieties datasets are divided into 80%/20%, training and test respectively (by default), for model evaluation. Another advantage of the SDM-CropProj framework is that it allows to set/control the number of modelling techniques employed, the number of pseudo-absences sets, and/or model rounds, as these factors strongly influence the computational time of the analyses. On the other hand, model accuracy can be assessed by different metrics, such as the Boyce index and Cohen's Kappa. However, two metrics are primarily used: i) the Area Under the Receiver Operating Characteristics (AUC-ROC) curve, a robust threshold-independent measure of a model's ability that yields the probability that a randomly selected presence will have a higher predicted value than a randomly selected absence (AUC ranges between 0 and 1, with measures below 0.7 are considered poor, 0.7–0.9 moderate, and $>$ 0.9 good) [14]; and, ii) the True Skill Statistic (TSS) maximum values of the ensemble models as a threshold-dependent measure of model accuracy (TSS ranges between -1 and 1, with measures below 0.4 are considered poor, 0.4–0.8 useful, and 0.8 good-excellent) [15]. Although there are underlying concerns related to both AUC or TSS, both are the only available discrimination metrics implemented in the *biomod2* package, which are amply used and highly suitable to test model performance.

The SDM-CropProj method also allows reducing the overall number of (partial) models in the final ensemble by selecting the best-performing ones. Given the large number of models usually generated per crop species/variety, the less performant models can be filtered out before the final ensemble forecasting. To do so, we can select a priori the top percentile best models (top-ranked models) for the best-performing techniques considering the AUC (or other metrics) distribution. Based on these top-performing models, an ensemble using the average value of all the partial projections is implemented, thus reducing inter-model uncertainty. Although the average is a straightforward measure to perform a multi-algorithm combination, there are other statistical metrics to quantify uncertainty between different methods (e.g., coefficient of variation, standard-errors, confidence intervals). Therefore, this approach allows to compare and assess predictions obtained from different methods/algorithms, each one with different characteristics, limitations, and advantages. By changing the input parameters

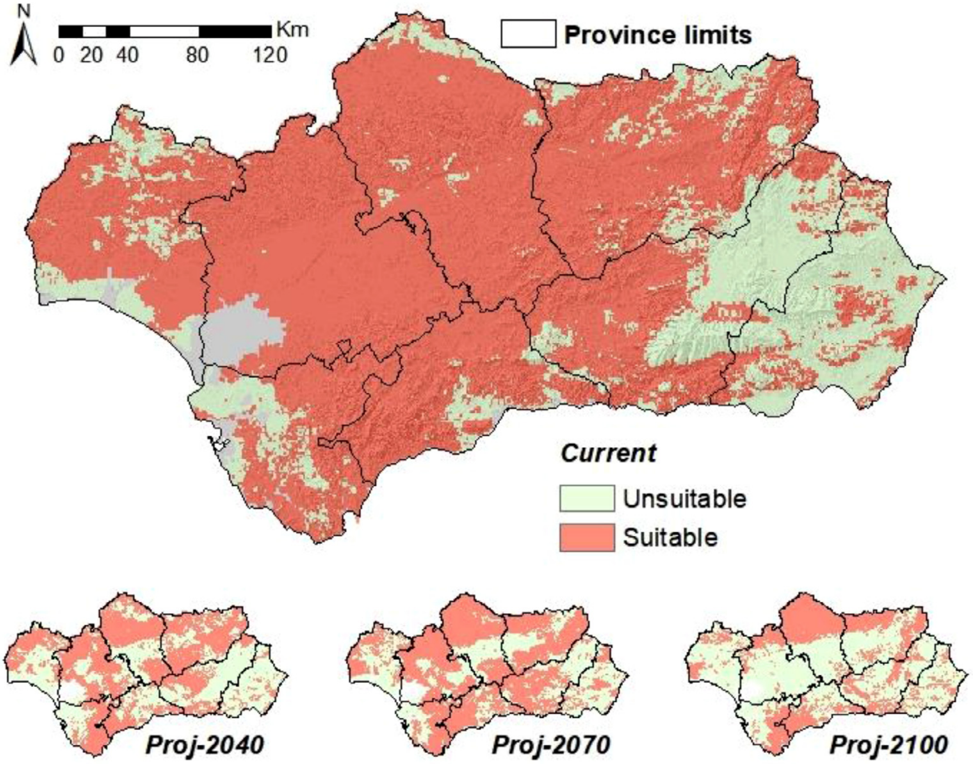

Fig. 1. Spatial projections of suitable and unsuitable areas for seven olive tree varieties derived from ensembled SDMs based on climate predictors for current and three future periods (2040, 2070, and 2100) in Andalusia (southern Spain). Source: https://github.com/salvador-arenas-castro/SDM-CropProj.git.

in SDM-cropProj, the modeller can combine partial projections in different ways and also quantify uncertainty measures. The suitability threshold value that minimized the distance between the AUC-ROC curve and the (0, 1) point is used for the binary transformations of model probability predictions into dichotomous suitable/unsuitable habitat maps. Finally, in our example, the ensemble model projections for future conditions are obtained by replacing the current (bio-)climatic predictors used in the calibration step with the regional climatic projections for the three future time points: 2040, 2070, 2100 (Fig. 1).

Fitting the log-log productivity model and evaluation

To analyze the relationship between the current area suitable for crops obtained in previous steps (transformed by log10) and the annual production (also transformed to log10 average production) by surface/region, the SDM-CropProj framework fits a straightforward univariate log-log linear model. A previous diagnostic of the regression model can be performed using the function ‘gg_diagnose’ in ‘lindia’ v0.9 R package (Fig. S1). Different metrics for assessing the predictive ability of the productivity model (e.g., the R^2 value, the root-mean-square error (RMSE), the Spearman correlation, and a leave-one-out cross-validation) can be used [4]. This step in the modeling process allows to relate crop production to model-predicted suitable area, and therefore future-suitability maps can be used to obtain projections of total fruit production. Furthermore, this method also allows for tracking potential trends of annual production due to changes in environmental suitability modulated by CC scenarios. This process is done by replacing the predictive/independent variable

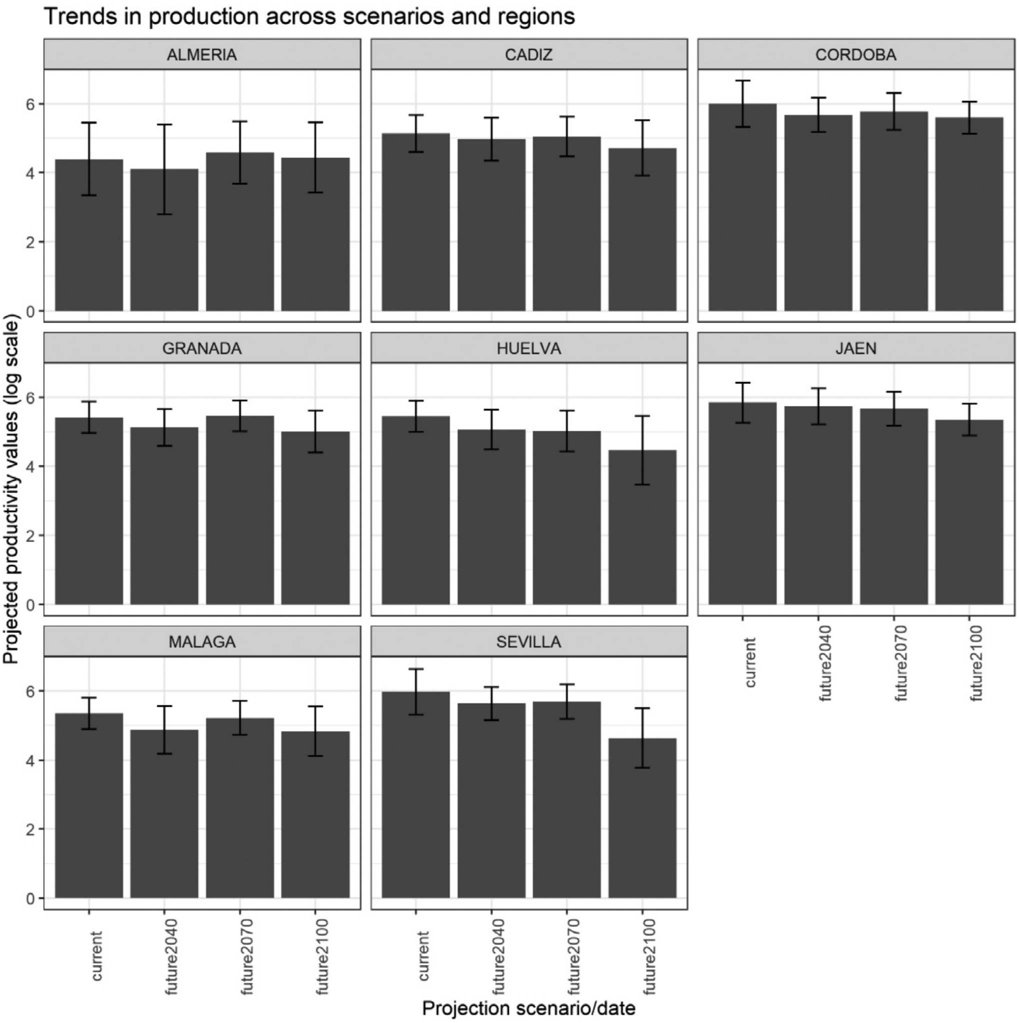

Fig. 2. Projected trends in fruit production (in log10 tons) for the main olive tree varieties cultivated in Andalusia (southern Spain) based on predicted suitable area by province and across different climate change scenarios. Bars represent the productivity predictions with standard error variance. Source: https://github.com/salvador-arenas-castro/SDM-CropProj.git.

(i.e., log10 area with suitable environmental conditions) with its projections from SDMs (i.e., step one; Fig. 2). In the proposed framework, changes in suitability are always motivated by climate scenarios since these constitute the dynamic part of the modelling system. Understanding which specific climatic factors/predictors are causing changes in suitable areas for each crop species/variety can be addressed by quantifying variable importance measures and response curves. These measures are easily obtained from ensemble models using the 'biomod2' R package and the SDM-cropProj framework (see example).

Method validation

Different methods of cross-validation of the SDM-CropProj framework have been described in previous sections.

Conclusion

The model-assisted SDM-CropProj framework allows forecasting the environmental suitability of crops and total annual production by region. Thus, this method provides an early-warning system potentially capable of detecting and anticipating spatiotemporal changes in spatial distribution and crop production in the face of climate change. In addition, it is expected to be applicable to a wide range of crop species or varieties and agroecosystems worldwide. Therefore, this straightforward and early-warning model-assisted framework that requires low-data amounts can become a valuable tool to support the current decision-making processes and optimize resources by stakeholders.

Acknowledgments

We sincerely thank Dr. Rafael Villar, who provided the needed resources to publish this article. The authors are especially grateful to Mariana Novo for using her talents to improve the illustrations. SAC was financially supported by PORBIOTA-E-Infraestrutura Portuguesa de Informação e Investigação em Biodiversidade (POCI-01-0145-568 FEDER-022127). JG was supported by FCT Scientific Employment Stimulus 2018 - CEECIND/02331/2017 and the support of the project proMetheus – Research Unit on Materials, Energy and Environment for Sustainability, FCT Ref. UID/05975/2020, financed by national funds through the FCT/MCTES.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.mex.2021.101394.

References

[1] J. Elith, J.R. Leathwick, Species distribution models: ecological explanation and prediction across space and time, Annu. Rev. Ecol. Evol. Syst. 40 (2009) 677–697, doi:10.1146/annurev.ecolsys.110308.120159.

[2] L.D. Estes, B.A. Bradley, H. Beukes, D.G. Hole, M. Lau, M.G. Oppenheimer, R. Schulze, M.A. Tadross, W.R. Turner, Comparing mechanistic and empirical model projections of crop suitability and productivity: implications for ecological forecasting, Glob. Ecol. Biogeogr. 22 (2013) 1007–1018, doi:10.1111/geb.12034.

[3] G. Besnard, B. Khadari, M. Navascués, M. Fernández-Mazuecos, A. El Bakkali, N. Arrigo, D. Baali-Cherif, V. Brunini-Bronzini de Caraffa, S. Santoni, P. Vargas, V. Savolainen, The complex history of the olive tree: from Late Quaternary diversification of Mediterranean lineages to primary domestication in the northern Levant, Proc. R. Soc. B Biol. Sci. 280 (2013) 20122833, doi:10.1098/rspb.2012.2833.

[4] S. Arenas-Castro, J.F. Gonçalves, M. Moreno, R. Villar, Projected climate changes are expected to decrease the suitability and production of olive varieties in southern Spain, Sci. Total Environ. 709 (2020) 136161, doi:10.1016/j.scitotenv.2019.136161.

[5] N.A.C. Cressie, Statistics for Spatial Data, John Wiley & Sons, Inc., Hoboken, NJ, USA, 1993, doi:10.1002/9781119115151.

[6] R.J. Hijmans, Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model, Ecology 93 (3) (2012) 679–688, doi:10.1890/11-0826.1.

[7] G.S. Cumming, Using between-model comparisons to fine-tune linear models of species ranges, J. Biogeogr. 27 (2000) 441–455, doi:10.1046/j.1365-2699.2000.00408.x.

[8] F.E.Jr. Harrell, K.L. Lee, R.M. Califf, D.B. Pryor, R.A. Rosati, Regression modelling strategies for improved prognostic prediction, Stat Med 3 (1984) 143–152, doi:10.1002/sim.4780030207.

[9] J. Franklin, Mapping species distributions, Cambridge University Press, Cambridge, 2010, doi:10.1017/CBO9780511810602.

[10] T.H. Booth, H.A. Nix, J.R. Busby, M.F. Hutchinson, <scp>bioclim</scp>: the first species distribution modelling package, its early applications and relevance to most current <scp>MaxEnt</scp>studies, Divers. Distrib. 20 (2014) 1–9, doi:10.1111/ddi.12144.

[11] X. Feng, D.S. Park, Y. Liang, R. Pandey, M. Papeş, Collinearity in ecological niche modeling: confusions and challenges, Ecol. Evol. 9 (2019) 10365–10376, doi:10.1002/ece3.5555.

[12] REDIAM, El clima de andalucía en el siglo XXI, Escenarios Locales de Cambio Climático de Andalucía (2014) https://bit.ly/2zC01WN.

[13] W. Thuiller, D. Georges, R. Engler, Biomod2: Ensemble Platform for Species Distribution Modeling, R Package Version 3.1-48, 2014 https://bit.ly/3gwnZ6w.

[14] J.M. Lobo, More complex distribution models or more representative data? Biodivers. Informatics. 5 (2008), doi:10.17161/bi.v5i0.40.

[15] O. Allouche, A. Tsoar, R. Kadmon, Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS), J. Appl. Ecol. 43 (2006) 1223–1232, doi:10.1111/j.1365-2664.2006.01214.x.