Species distribution modelling of the genus *Equisetum* subgenus *Equisetum* for the territory of Russia

D. S. Feoktistov[*], E. Zh. Baiakhmetov

National Research Tomsk State University, Lenina ave. 36, 634050, Tomsk, Russia,
*Corresponding author E-mail: feoktistovdmitriy@gmail.com

Horsetails are a complex taxonomic and systematic group. Therefore, the study of the geographical distribution of these species is necessary for a better understanding of the phylogeny of this family. We concluded an analysis of the distribution of 5 species of horsetail of the subgenus *Equisetum* (*Equisetum*, Equisetaceae): *E. arvense* L., *E. fluviatile* L., *E. palustre* L., *E. pratense* Ehrh., *E. sylvaticum* L. using the maximum entropy method implemented in the MaxEnt program. Modeling was carried out using climate variables from the WorldClim global climate base. Simulation results show good simulation quality. In 3 out of 5 species, the AUC of the test sample was in the range of 0.9–1, and in 2 species — 0.8–0.9. In general, for most species, a plausible picture of their intended distribution has developed. The obtained models suggest that the territory of Russia is favorable enough for the growth of horsetails. Analysis of the contribution of 14 bioclimatic variables to the distribution of the studied species revealed that the most important variables are: annual mean temperature, isotermality, temperature seasonality, max temperature of warmest month, temperature annual range, mean temperature of warmest quarter, mean temperature of driest quarter, mean temperature of coldest quarter, annual precipitation, precipitation of wettest month, precipitation seasonality, precipitation of driest quarter, precipitation of warmest quarter, and precipitation of coldest quarter.

Key words: Equisetum; species distribution modeling; siberia; maxEnt

Introduction

The genus *Equisetum* L. (Equisetaceae) is represented by a small number of living species. According to the currently accepted pteridophyte systems, it is divided into 2 subgenus *Equisetum* and *Hippochaete* (Milde) Baker (Hauke, 1963; 1978), which are sometimes understood as independent genera Equisetum s.str. and Hippochaete Milde (Farwell, 1916; Rothmaler, 1944). They are very complex group in terms of taxonomy. Horsetails are widely distributed in temperate latitudes. Studying geographical distribution of horsetails are important to fully understand biodiversity of the region. Therefore, in the present study we conducted species distribution modelling analysis on *Equisetum* species. This paper focuses on subgenus *Equisetum* and five species that is commonly found in the area: *E. arvense* L., *E. fluviatile* L., *E. palustre* L., *E. pratense* Ehrh., *E. sylvaticum* L.

Materials and methods

The purpose of the SDM is to assess the similarity of conditions in the study area with conditions in places of a known occurrence (and, in some cases, the absence of this phenomenon). Basically, this method is used to predict the geographical distribution of species using climate data as distribution factors. The analysis consists of the following steps: 1) places of occurrence of species are noted; 2) the values of variable ecological predictors (for example, climate) are taken on the studied territories, they can be obtained from geographic information databases; 3) environmental criteria are used in creating a model for assessing similarities with the places of findings, or other values, such as the abundance of species; 4) the model is used to predict the distribution of interest on the scale of the entire area of interest (and, possibly, for the future or past climate) (Hijmans & Elith, 2013). There are a number of approaches for implementing this method, using the statistical software package R (Hijmans & Elith, 2013), or GIS (geographic information system) ArcGIS 10.2 and special scripts for it (SDM Toolbox) (Brown, 2014), which facilitate the preparation of data for analysis, since it is the preparation of the data that takes the main time for the analysis. The data obtained on the distribution of species reflect an equilibrium relationship to a given environment.

Among the many algorithms for modeling the spatial distribution of species (Broennimann et al., 2007; Guisan & Thuiller, 2007; Predictive habitat distribution models in ecology, 2000; Guo & Liu, 2010; Stigall, 2012; Stockwell, 1999), the most popular method is the maximum entropy implemented in the MaxEnt program (Maximum Entropy Species Distribution Modeling) (Elith et al., 2011; Phillips et al., 2006; Phillips & Dudík, 2008). The initial data for modeling in this case are only the presence of the species (presence-only data). The principle of this algorithm is as follows: prediction must satisfy all the limitations known to the researcher, and the desired distribution must have maximum entropy (Phillips et al., 2006). The result of the algorithm is a habitat suitability model — a map with predicted probabilities of the presence of a species in each raster cell. In the science of vegetation, modeling the distribution of species, in particular, using MaxEnt, is used to solve environmental problems, identify the distribution of invasive species and predict their further distribution (Dobrowski et al., 2011; Peterson, 2005; Smith et al., 2008), comparing niches of hybrids and their parent species (Engler et al., 2013), revealing the distribution parameters of species richness (Cord et al., 2014).

An analysis of the spatial distribution of horsetail species was carried out for the territory of Russia, since the coordinates of the points of occurrence were obtained for this territory. To carry out the analysis, we took the following steps.

1. Data collection. To compile the species distribution database, we used data from herbarium samples, various flora, as well as data from GBIF (Global Biodiversity Information Facility). In the case of determinants and flora, when data on the distribution of the species are presented in the form of a map, the map was georeferenced in the ArcGIS program and the coordinates of the marked points were determined in the coordinate system WGS1984. This step is necessary due to the fact that maps from different sources can have different projections, and accordingly different coordinates. Particular attention was paid to the coordinates of the habitats of the samples collected by us during the field trips.

2. Bioclimatic variables were downloaded from wordlclim.org (Table 1). They are most often used in such studies. A total of 19 bioclimatic variables were loaded. Bioclimatic variables have a resolution of 30 seconds, which is 1 km^2 near the equator. Then, we reduced the highly correlated climate variables, since using them in the model can contribute to the overfitting of the model. To do this, using the SDM Toolbox, we analyzed the correlation of climatic variables that will be used to build the model. At the output, this tool provides a matrix of distances at which bioclimatic variables are indicated and the correlation coefficient between them. Bioclimatic variables with a correlation coefficient greater than 0.7 were not used to build a further model.

Table 1. Bioclimate variables. Highly correlated variables with k <0.7 are highlighted by bold.

Code	Description
BIO1	**Annual Mean Temperature**
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3	**Isothermality (BIO2/BIO7) (* 100)**
BIO4	**Temperature Seasonality (standard deviation ×100)**
BIO5	**Max Temperature of Warmest Month**
BIO6	Min Temperature of Coldest Month
BIO7	**Temperature Annual Range (BIO5-BIO6)**
BIO8	Mean Temperature of Wettest Quarter
BIO9	**Mean Temperature of Driest Quarter**
BIO10	**Mean Temperature of Warmest Quarter**
BIO11	**Mean Temperature of Coldest Quarter**
BIO12	**Annual Precipitation**
BIO13	**Precipitation of Wettest Month**
BIO14	Precipitation of Driest Month
BIO15	**Precipitation Seasonality (Coefficient of Variation)**
BIO16	Precipitation of Wettest Quarter
BIO17	**Precipitation of Driest Quarter**
BIO18	**Precipitation of Warmest Quarter**
BIO19	**Precipitation of Coldest Quarter**

Next, climate heterogeneity of the studied region was analyzed. This is one of the steps for spatial rarification of data. To analyze the climate heterogeneity, the Principal Component Analysis was used, which is implemented in the SDM Toolbox. The method of principal components extracts the principal ones from all climatic variables and then for each raster point calculates the value to which a certain color is assigned. The first component accounts for 87.7% of the variability, the second 10.2%, the third 2%.

In order for the model to produce data that corresponds with real distribution, most SDM methods require the input of data from spatially independent collection points. The elimination of spatial clusters of occurances is essential for calibrating and evaluating the model. If points of occurrence are presented more often in a certain territory (for example, roadsides or neighborhoods of cities), the model shows more suitable environmental conditions in these places (this reduces the model's ability to predict spatially independent data) and the model's characteristics may be overestimated (Boria et al., 2014; Hijmans, 2012; Veloz, 2009). The tools included in the SDM Toolbox allow users to spatially thin out their data at several levels depending on habitat, topographic heterogeneity or climate heterogeneity. For example, habitats can be filtered out for 5 km^2, 10 km^2 and 30 km^2 in areas with high, medium and low environmental heterogeneity, respectively. This filtering method is especially useful for studies with a limited number of points of occurrence points and can maximize the number of spatially independent observations. In our case, we filtered out the presence points at a distance of 25 km^2 in the case of areas with a similar climate, and 5 km^2 in cases with areas with a heterogeneous climate. Then we created a file for background points. It selects the background points and the density of the background sample.

Background points (and similar pseudo-absence points) are intended to be compared with presence data and help differentiate environmental conditions under which species can potentially grow. Typically, background points are selected within a large rectangular area where habitats often exist that are ecologically suitable but not yet colonized. When background points are selected in these habitats, this increases omission errors (false positives). As a result, the "best" model tends to be overly favorable and shows those habitats on the map where the species may not grow (Barbet-Massin et al., 2012).

3. Launch of MaxEnt (Version 3.3.3 e) (Phillips & Dudík, 2008) to create a distribution model. Each model was built in 20 replicates; during the simulation, the initial sample was randomly divided into training (75%) and test (25%) sets (repeated sampling method - subsample) (Phillips & Dudík, 2008).

The quality of the obtained models was studied both statistically and on the basis of expert ideas about the distribution of species in nature. For statistical evaluation, the AUC indicator (area under receiver operating characteristic (ROC) curve), a non-parametric hierarchical tool used to assess the predictive ability of the chosen model (Fielding & Bell, 1997). From the results obtained during the simulation, the best AUC calculated for the training and test data sets was selected according to the minimum standard deviation (Warren & Seifert, 2011).

According to the AUC indicator, modeling quality is usually divided into 5 categories (Swets, 1988): 0.9–1 = "excellent", 0.8–0.9 = "good", 0.7–0.8 = "satisfactory", 0.6–0.7 = "bad", <0.6 - "very bad" (modeling failed).

When building the model, 10 percentile thresholds was also considered, which allows to cut off points that are in the area in which the probability of finding the species is below 10%.

The contribution of environmental factors in the creation of the model was evaluated by the importance for permutation test (Phillips et al., 2006; Phillips & Dudík, 2008).

Results

Species distribution modelling results show good model quality. In 3 out of 5 species, the AUC of the training sample was in the range of 0.9–1, and in 2 species — 0.8–0.9. For the test sample, 3 species had excellent AUC, and 2 species had good AUC. In general, for most species, a plausible picture of their intended distribution has developed. To analyze the contribution of climate variables, let us dwell on the modeled distribution of species in more detail.

Subgenus *Equisetum*. *E. arvense* (Figure 1) demonstrated a model with one of the lowest AUC values among all other species. AUC for this species was 0.869 with a standard deviation of 0.012. Many researchers note that species with wide ecological amplitudes show worse modeling results (Guisan et al., 2007; Morán-Ordóñez et al., 2012; Stockwell & Peterson, 2002). For *E. arvense*, the following climate variables are most important (indicated in order of greatest contribution to model building): bio5 - maximum temperature of the warmest month, bio7 - temperature annual range, bio3 - isothermality, bio15 - precipitation seasonality and bio13 - precipitation of wettest month. Such a distribution of climatic variables suggests that air temperature is more important for this species than precipitation and humidity. However, variables 4 and 5 are responsible for the amount of precipitation during the season, which indicates the need for precipitation throughout the growing season. The resulting picture corresponds to environmental ideas about this species. The most suitable habitats for *E. arvense* were found in the boreal zone of Eurasia, especially in the central and northern parts of European Russia, in the south of Western and Central Siberia. Judging by the resulting model, *E. arvense* has a very wide ecological amplitude and can occupy a variety of ecological niches.
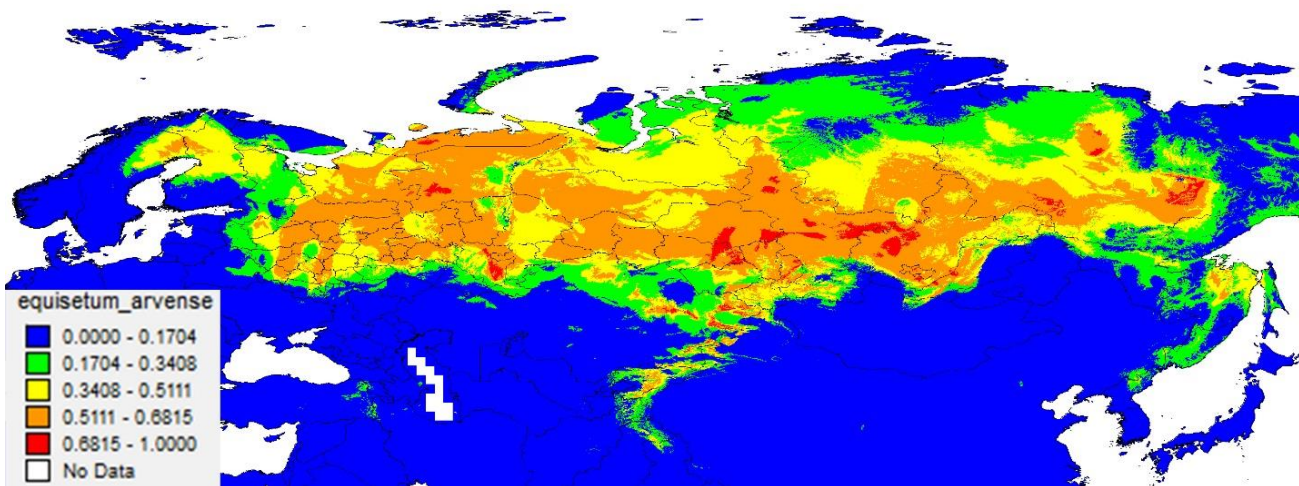

Figure 1. Map of *E. arvense* distribution model, obtained by method of maximum enthropy in MaxEnt. Visualised in DivaGIS. Warmer colours indicate more suitable habitats.

E. fluviatile (Figure 2). AUC for this species showed excellent modelling quality. The AUC value for this species is 0.921 with a standard deviation of 0.020. For *E. fluviatile*, the following climate variables are most important (indicated in order of greatest contribution to model building): bio7 - temperature annual range, bio1 - annual mean temperature, bio18 - precipitation of warmest quarter, bio3 - isothermality and bio10 - mean temperature the warmest quarter. Such a distribution of climatic variables suggests that temperature throughout the year is of greater importance for this species; variables 3 and 5 are responsible for the amount of precipitation during the warmest quarter, which indicates the need for precipitation during the growing season. The resulting picture corresponds to environmental ideas about this species. The most suitable habitats for *E. fluviatile* were found in the northern part of European Russia. In Western Siberia and the Urals, less suitable growing conditions for this species. Within Siberia, the most suitable conditions for this species are observed in Altai.
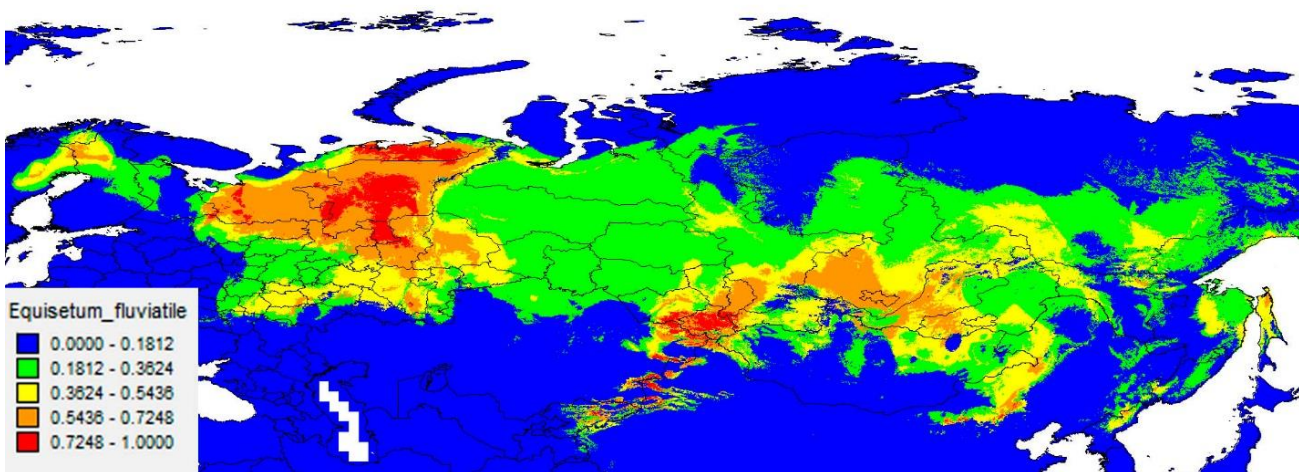

Figure 2. Map of *E. fluviatile* distribution model, obtained by method of maximum enthropy in MaxEnt. Visualised in DivaGIS. Warmer colours indicate more suitable habitats.

E. palustre (Figure 3). AUC for this species was excellent and amounted to 0.929 with a standard deviation of 0.015. For *E. palustre*, the most important climatic variables are (indicated in the order of contribution to the model): bio1 - annual mean temperature, bio18 - precipitation of warmest quarter, bio3 - isothermality, bio12 - annual precipitation and bio17 - precipitation of driest quarter. Such a distribution of climatic variables suggests that the annual mean temperature in the growing area is important for this specie. The remaining variables are responsible for the amount of precipitation per quarter and the mean annual precipitation, which indicates a significant dependence of this species on precipitation and humidity. The resulting picture corresponds to environmental ideas about this species. The distribution map of *E. palustre* was very close to that of *E. fluviatile*. This was probably due to the fact that these two species have similar environmental requirements for growing conditions. The most suitable habitats for *E. palustre* were found in the central and even northern parts of European Russia. Outside the Urals, the most favorable conditions are formed in the extreme south-west of Western Siberia and Altai; somewhat less favorable conditions are found in the mountains of Southern Siberia. The limiting factor, in all likelihood, is temperature for him, since he is not in the high latitudes of Siberia. The method did not show the suitability of the conditions of the central and southern parts of European Russia due to the insufficient number of points from this territory.
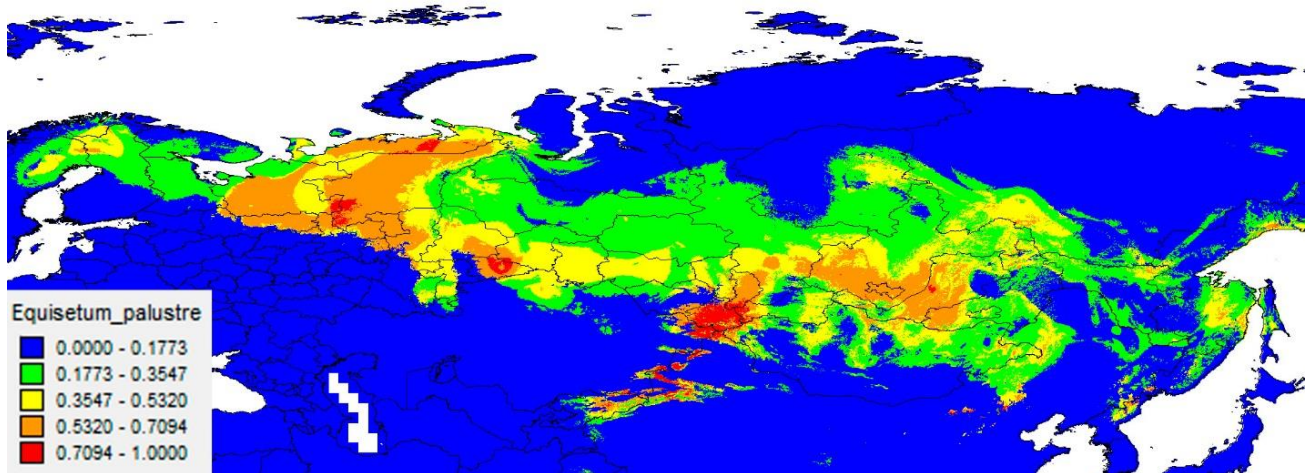

Figure 3. Map of *E. palustre* distribution model, obtained by method of maximum enthropy in MaxEnt. Visualised in DivaGIS. Warmer colours indicate more suitable habitats.

E. pratense (Figure 4). AUC for this species showed the smallest value of all modeled species was 0.823 with a standard deviation of 0.025. But according to the conditional classification of models, it can still be attributed to good ones. *E. pratense*, like *E. arvense*, has high ecological plasticity, which is possible and is the reason for less successful modeling. For *E. pratense*, the most important climatic variables are (indicated in the order of contribution to the model): bio9 – mean temperature of driest quarter, bio1 - annual mean temperature, bio12 - annual precipitation, bio19 – precipitation of coldest quarter and bio17 - precipitation of the driest quarter. The first two climatic variables indicate that temperature plays the most important role for this specie. The total amount of precipitation and the amount of precipitation in the driest and coldest months is of no small importance, since this species often grows not on the banks of rivers or streams, where there is a lot of soil moisture, but in forests and meadows. The resulting model corresponds to the geographical and environmental ideas about the distribution of this species. The most favorable conditions for him turned out to be in Altai, in the region of the southern tip of Lake Baikal and in the south of the Far East, and somewhat less favorable in the Urals and in Southern Siberia.
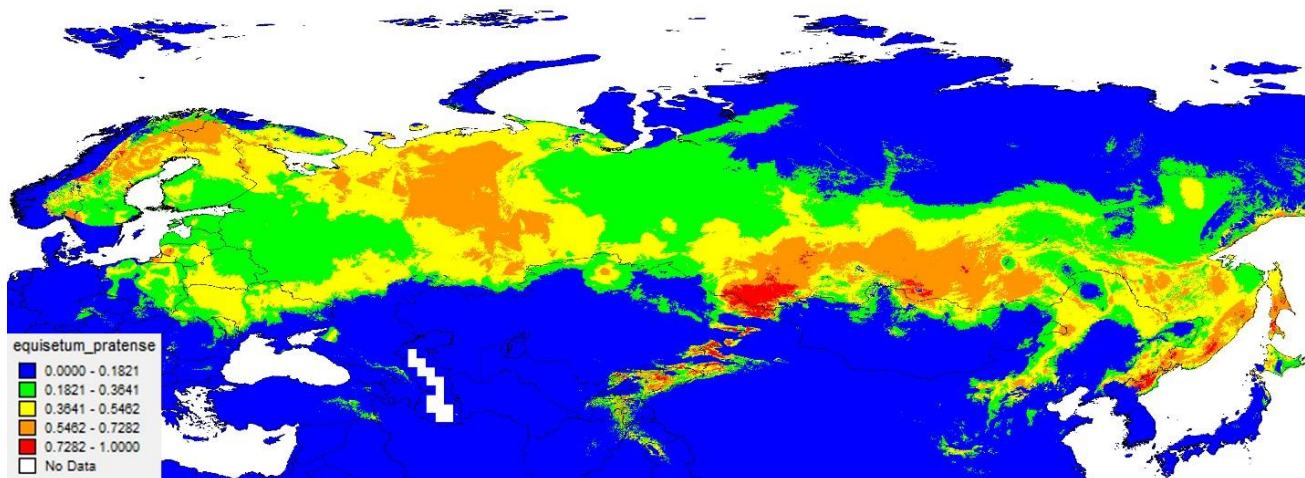

Figure 4. Map of *E. pratense* distribution model, obtained by method of maximum enthropy in MaxEnt. Visualised in DivaGIS. Warmer colours indicate more suitable habitats.

E. sylvaticum. The AUC for this species was 0.905 (excellent) with a standard deviation of 0.029. For *E. sylvaticum*, the most important climatic variables are (indicated in the order of contribution to the model): bio3 - isothermality, bio18 - precipitation of

warmest quarter, bio4 - temperature seasonality, bio1 – mean annual temperature, bio7 - temperature annual range. The distribution of this species is mainly due to climatic and temperature variables and, to a lesser extent, the effect of precipitation. The resulting model corresponds to the geographical and environmental ideas about the distribution of this species. The most suitable places for growth for it are the Southern and Middle Urals, and Altai, the Urals and Trans-Urals, and the southern part of Western and Central Siberia are somewhat less suitable.

Conclusion

For most species of horsetail, the territory of Russia is quite favorable for growth, as evidenced by the model of distribution of species obtained by the method of maximum entropy. An analysis of the contribution of 14 bioclimatic variables to the distribution of the studied species revealed that the most important variables are: annual mean temperature, isotermality, temperature seasonality, max temperature of warmest month, temperature annual range, mean temperature of warmest quarter, mean temperature of driest quarter, mean temperature of coldest quarter, annual precipitation, precipitation of wettest month, precipitation seasonality, precipitation of driest quarter, precipitation of warmest quarter, precipitation of coldest quarter.

Acknowledgments

References

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? Methods in Ecology and Evolution, 3(2), 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

Boria, R. A., Olson, L. E., Goodman, S. M., & Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. Ecological Modelling, 275, 73–77. https://doi.org/10.1016/j.ecolmodel.2013.12.012

Broennimann, O., Treier, U. A., Müller-Schärer, H., Thuiller, W., Peterson, A. T., & Guisan, A. (2007). Evidence of climatic niche shift during biological invasion. Ecology Letters, 10(8), 701–709. https://doi.org/10.1111/j.1461-0248.2007.01060.x

Brown, J. L. (2014). SDMtoolbox: A python-based GIS toolkit for landscape genetic, biogeographic and species distribution model analyses. Methods in Ecology and Evolution, 5(7), 694–700. https://doi.org/10.1111/2041-210X.12200

Cord, A. F., Klein, D., Gernandt, D. S., de la Rosa, J. A., & Dech, S. (2014). Remote sensing data can improve predictions of species richness by stacked species distribution models: a case study for Mexican pines. Journal of Biogeography, 41(4), 736–748. https://doi.org/10.1111/jbi.12225

Dobrowski, S. Z., Thorne, J. H., Greenberg, J. A., Safford, H. D., Mynsberge, A. R., Crimmins, S. M., & Swanson, A. K. (2011). Modeling plant ranges over 75 years of climate change in California, USA: Temporal transferability and species traits. Ecological Monographs, 81(2), 241–257. https://doi.org/10.1890/10-1325.1

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. Diversity and Distributions, 17(1), 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

Engler, J. O., Rödder, D., Elle, O., Hochkirch, A., & Secondi, J. (2013). Species distribution models contribute to determine the effect of climate and interspecific interactions in moving hybrid zones. Journal of Evolutionary Biology, 26(11), 2487–2496. https://doi.org/10.1111/jeb.12244

Farwell, O. A. (1916). The genus Hippochaete in North America, north of Mexico. Mem NY Bot Gard, 6, 461–472.

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence / absence models. Environmental Conservation, 24(1), 38–49. https://doi.org/10.1017/S0376892997000088

Guisan, A., Zimmermann, N. E., Elith, J., Graham, C. H., Phillips, S., & Peterson, A. T. (2007). What matters for predicting the occurrences of trees: Techniques, data, or species' characteristics? Ecological Monographs, 77(4), 615–630. https://doi.org/10.1890/06-1060.1

Guisan, A, & Thuiller, W. (2007). Predicting species distribution: offering more than simple habitat models (vol 8, pg 993, 2005). Ecology Letters, 10(5), 435. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Predictive habitat distribution models in ecology, 135 Ecological Modelling 147 (2000). https://doi.org/10.1016/S0304-3800(00)00354-9

Guo, Q., & Liu, Y. (2010). ModEco : an integrated software package for ecological niche modeling. Ecography, March, 1–6. https://doi.org/10.1111/j.1600-0587.2010.06416.x

Hauke, R. L. (1963). A Taxonomic Monograph of the Genus Equisetum Subgenus Hippochaete. Beihefte Zur Nova Hedwigia, 8, 1–123. http://kbd.kew.org/kbd/detailedresult.do?id=25673

Hauke, R. L. (1978). A taxonomic monograph of Equisetum subgenus Equisetum. Nova Hedwigia, 385–456.

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

Hijmans, R. J., & Elith, J. (2013). Species distribution modeling with R Introduction. In October. https://doi.org/10.1016/S0550-3213(02)00216-X

Morán-Ordóñez, A., Suárez-Seoane, S., Elith, J., Calvo, L., & de Luis, E. (2012). Satellite surface reflectance improves habitat distribution mapping: A case study on heath and shrub formations in the Cantabrian Mountains (NW Spain). Diversity and Distributions, 18(6), 588–602. https://doi.org/10.1111/j.1472-4642.2011.00855.x

Peterson, E. B. (2005). Estimating cover of an invasive grass (Brontus tectorum) using tobit regression and phenology derived from two dates of Landsat ETM + data. International Journal of Remote Sensing, 26(12), 2491–2507. http://www.tandfonline.com/doi/abs/10.1080/01431160500127815

Phillips, S. B., P. Anderson, R., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. Ecological Modelling, 190(3–4), 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation. Ecography, 31(2), 161–175. https://doi.org/10.1111/j.0906-7590.2008.5203.x

Rothmaler, W. H. P. (1944). Pteridophyten studien. I. In Feddes Repertorium (Vol. 54, pp. 55–82). http://onlinelibrary.wiley.com/doi/10.1002/fedr.19440540106/abstract

Smith, M., Oliveira-Filho, A. T., Bachman., S., Moat, J., Lughada, E. M. N., & Lucas, E. J. (2008). Plant Diversity Hotspost in the atlantic Coastal Forests of Brazil. Conservation, 23(1), 151–163. http://onlinelibrary.wiley.com/doi/10.1111/j.1523-1739.2008.01075.x/full

Stigall, A. (2012). Using ecological niche modeling to evaluate niche stability in deep time. Journal of Biogeography, 39, 772–781. http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2699.2011.02651.x/full

Stockwell, D. (1999). The GARP modelling system: problems and solutions to automated spatial prediction. International Journal of Geographical Information Science, 13(2), 143–158. https://doi.org/10.1080/136588199241391

Stockwell, D. R. ., & Peterson, A. T. (2002). Effects of sample size on accuracy of species distribution models. Ecological Modelling, 148(1), 1–13. https://doi.org/10.1016/S0304-3800(01)00388-X

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. Science (New York, N.Y.), 240(4857), 1285–1293. https://doi.org/10.1126/science.3287615

Veloz, S. D. (2009). Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. Journal of Biogeography, 36(12), 2290–2299. https://doi.org/10.1111/j.1365-2699.2009.02174.x

Warren, D. L., & Seifert, S. N. (2011). Ecological niche modeling in Maxent: The importance of model complexity and the performance of model selection criteria. Ecological Applications, 21(2), 335–342. https://doi.org/10.1890/10-1171.1