Master Thesis
im Rahmen des
Universitätslehrganges „Geographical Information Science & Systems“
(UNIGIS MSc) am Interfakultären Fachbereich für GeoInformatik (Z_GIS)
der Paris Lodron-Universität Salzburg

zum Thema

„Utilization of two approaches of species distribution modelling (GLM, MaxEnt) to find new sites for seed collecting“ An application on six alpine vascular plant species in the eastern Alps (Styria)

vorgelegt von

Mag. Patrick Schwager
102834, UNIGIS MSc Jahrgang 2014

Zur Erlangung des Grades
„Master of Science (Geographical Information Science & Systems) – MSc(GIS)”

Gutachterin
Dr. Gudrun Wallentin

Gratwein-Straßengel, 26.05.2017

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Master ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Diese Arbeit wurde in gleicher oder ähnlicher Form noch bei keiner anderen Prüferin/ keinem anderen Prüfer als Prüfungsleistung eingereicht.

Die vorgelegte Fassung entspricht der eingereichten elektronischen Version.

Acknowledgements

I would especially like to thank my wife Julia and my son Paul, who often had cope with my absence as my studies came to an end. Now better times are on the way!

I would like to thank the entire UNIGIS team for the excellent student support and the pleasant climate.

I would like to thank Dr. Gudrun Wallentin for the supervision and for the patience in terms of timing.

I would also like to thank my colleagues at the University of Graz.

My thanks also go to my boss, Dr. Christian Berg, who has made it possible for me to implement the work as part of our project. He also assisted with species photos as his were much clearer than the ones that I took myself.

Thanks also to Dr. Viktoria Wagner for her support in statistical questions.

Abstract

Since 2015, the Botanical Garden Graz has been part of the Alpine Seed Conservation and Research Network initialized by the Millennium Seedbank (Botanical Gardens, Kew). The overarching aim of the project is to use the European Alpine Seed Conservation Network to improve the conservation status of endangered plant species and communities in their habitats in the European Alps. Therefore, all project partners agreed to collect seeds from at least 100 vascular plant species from different regions of the Alps to reach the goal of 500 species for ex-situ conservation.

Using the example of six species distribution models from alpine vascular plants, this thesis investigates whether models can be helpful in finding suitable collection areas. Two different and widely used approaches were compared; the proven generalized linear model (GLM) and the machine learning algorithm MaxEnt.

Both modelling approaches were able to predict the distribution of the six vascular plant species across the Styrian Alps. The models were able to make plausible predictions and comply with known distributions provided by the Styrian distribution atlas. The prediction maps of both approaches show very similar results for one species, whereby GLM models tend to make less restrictive estimations. This means that the latter method deemed more regions as suitable for collection.

With regard to the Alpine Seed Conservation and Research Network, species distribution models can assist in localizing special areas of interest. With the help of the prediction maps it is possible to restrict field surveys to particular areas that show higher probability values. And this could contribute significantly to the collection success.

Zusammenfassung

Seit 2015 ist der Botanische Garten Graz Teil des Alpine Seed Conservation and Research Network, das von der Millennium Seedbank (Botanical Gardens, Kew) initialisiert wurde. Das übergeordnete Ziel des Projektes ist es, das europäische Alpine Seed Conservation Network zu nutzen, um den Erhaltungszustand von gefährdeten Pflanzenarten und Pflanzengesellschaften in ihren Lebensräumen in den europäischen Alpen zu verbessern. Deshalb haben sich alle Projektpartner darauf geeinigt, Saatgut von mindestens 100 Gefäßpflanzenarten aus verschiedenen Alpenregionen zu sammeln, um das gemeinsame Ziel von 500 Arten für die ex-situ-Erhaltung zu erreichen.

Am Beispiel von sechs Artverteilungsmodellen alpiner Gefäßpflanzen untersucht diese Arbeit, ob Modellierungen bei der Suche nach geeigneten Sammelgebieten hilfreich sein können. Dazu wurden zwei verschiedene und weit verbreitete Ansätze verglichen; Das bewährte generalisierte lineare Modell (GLM) und der machine-learning Algorithmus MaxEnt.

Beide Modellierungsansätze konnten die Verteilung der sechs Gefäßpflanzenarten in den steirischen Alpen vorhersagen. Die Modelle konnten plausible Vorhersagen machen und entsprechen weitgehend den Verbreitungsangaben aus dem Verbreitungsatlas der steirischen Farn- und Blütenpflanzen. Die Prognosekarten beider Ansätze zeigen für ein und dieselbe Art sehr ähnliche Ergebnisse, wobei GLM-Modelle dazu neigen, weniger restriktiv zu schätzen. Das heißt, die Methode identifiziert mehr Regionen als geeignete Gebiete.

Im Hinblick auf das Alpine Seed Conservation and Research Network können Artverbreitungsmodellierungen dazu beitragen, spezielle Gebiete zu lokalisieren. Mit Hilfe der Prognosekarten ist es möglich, Sammelexkursionen auf bestimmte Bereiche zu beschränken, welche höhere Vorkommenswahrscheinlichkeiten aufweisen. Und das könnte wesentlich zum Sammlungserfolg beitragen.

List of figures

List of tables

List of abbreviations

AC	Autocovariate
AIC	Akaike Information Criterion
AUC	Area under the curve
CBD	Convention on Biological Diversity
DEM	Digital elevation model
e. g.	exempli gratia
GEE	Generalized estimating equation
GIS	Geographical Information System
GLM	Generalized linear model
GLMM	Generalized Llinear mixed model
GSPC	Global Strategy of Plant Conservation
GWR	Geographically weighted regression
i. a.	inter alia
IUCN	International Union for Conservation of Nature
MaxEnt	Maximum entropy (Algorithm)
MSB	Millennium Seed Bank
NA	not available (missing Data)
PO	presence-only
ROC	Receiver operating characteristic
SAC	Spatial autocorrelation

SCI	Site of community importance
SDM	Species distribution model
SEVM	Spatial eigenvector mapping
VIF	Variance inflation factor

Content

1 INTRODUCTION

The vascular plant flora of the Alps has never been so endangered nor has it been as insufficiently protected as it is today. In addition to anthropogenic interferences, global warming is one of the main threats to the alpine flora. However, the Alps are of particular importance due to their endemic wealth. High levels of endemism are found by Tribsch (2004) in the southern, southwestern, easternmost and north eastern Eastern Alps. Above the tree line, zonal alpine grassland and azonal vegetation (screes, rocks, snowbeds) are essential to the endemic flora and a clear preference was observed for calcareous bedrock (Essl *et al.* 2009).

The Global Strategy of Plant Conservation (GSPC) is a legally binding component of the Convention on Biological Diversity (CBD). The long-term objective of the 16-point-program is to preserve the current and continuing loss of plant diversity (Jackson & Kennedy 2009)

The preferred method of species conservation is in-situ conservation, enabling species to fulfil their ecological roles in an optimal manner (Oldfield 2009). On the other hand, ex-situ conservation is of particular importance in addition to in-situ measures. Item 8 of the GSPC is of crucial relevance for botanical gardens, herbaria and botanical institutions. It is formulated with the intention of securing at least 75 % of threatened plant species in ex situ collections, preferably in the countries of origin, and to make at least 20 % available for recovery and restoration programmes.

It was for those reasons that the Botanical Garden of Graz (GZU, Institute for plant science, Karl-Franzens-University) established the first long-term seed bank for wild vascular plants growing within the Austrian province of Styria (Steiermark) in 2008. The project’s main objective is to collect and store diaspores from all of Styria’s wild vascular plants including herbarium voucher specimens, accompanied by location and habitat data (Gosch & Berg 2010).

Since 2013, the Botanical Garden of Graz has been collaborating with the Millennium Seed Bank (MSB, Royal Botanic Gardens, Kew) on several projects and our activities have taken place within an international context. Since 2015, the Botanical Garden Graz has been part of the Alpine Seed Conservation and Research Network (Müller *et al.* 2017). The overarching aim of the project is to use the European Alpine Seed Conservation Network to improve the conservation status of endangered plant species and communities in their habitats in the European Alps. Therefore all project partners agreed to collect seeds from at least 100 vascular plant species from different regions of the Alps to reach the goal of 500 species for ex-situ conservation. This contributes to the aim of safeguarding seeds from 25 % of the world flora by 2020, in a network of seed banks throughout the Millennium Seed Bank Partnership.

1.1 SPECIES DISTRIBUTION MODELS IN CONTEXT OF NATURE CONSERVATION

In order to protect species it is important to understand the connections between species and their abiotic and biotic environment. Species distribution Models (SDM), also called ecological niche models or habitat suitability models, utilize relationships between environmental variables and species observations to find environmental conditions where these populations could potentially occur. SDMs are beneficial for ecologists with regard to wildlife and resource management, conservation management or restoration ecology. These models are often used to locate occurrences of species or to identify the areas that are most important for conservation actions. Parolo et al. (2008) modelled the niche of *Arnica montana* in a Site of Community Importance (SCI) in the Italian Alps (Val Viola Bormina-Ghiacciaio di Cima dei Piazzi). The work contributes to the conservation management plan of the Natura 2000 area. In a project similar to the Alpine Seed Conservation and Research Network, they collected seeds for ex-situ conservation in collaboration with the Millennium Seed Bank. Williams et al. (2009) identified SDM approaches using presence-only data, like random forest and MaxEnt, as effective tools for discovering new populations of rare plant species in the Shasta-

Trinity National Forest in northern California. They also realized that a maximum observation number is necessary to achieve suitable models.

SDMs are based on ecological or evolutionary theories and thus provide a practical framework for answering questions concerning applied ecology or conservation biology and are extremely relevant to fundamental science like biogeography and phylogeography (Guisan & Thuiller 2005) as well as to environmental and climate change. Sérgio et al. (2007) show the importance of geographical range in the context of evaluating IUCN status of endangerment by modelling four bryophyte species. Lomba et al. (2010) overcome the “rare species modelling paradox” and provide a solution that includes a large number of predictors ensuring that the models are not over fitted. The proposed modelling framework provides a basis for adaptive conservation, management and monitoring of rare species at distinct spatio-temporal scales.

In times of rapidly changing environmental conditions, SDM is gaining in importance. In order to forecast the climate-driven spatio-temporal dynamic of high mountain plants Dullinger et al. (2012) utilized static geographic projections of species’ habitat shift in combination with demography simulations and seed dispersal in a hybrid model. The results are alarming as they predict a loss of habitat range size of 44-50 % by the end of the twenty-first century. Endemic species are the ones most affected by this habitat loss. A dynamic eco-evolutionary forecasting framework with regard to climate change was recently presented by Cotto et al. (2017) for four endemic perennial plant species in the Austrian Alps. The models combine niche modelling with individual-based demographic and genetic simulations and demonstrate how perennial species persist in unsuited habitats (due to climate change) longer than predicted by niche modelling.

SDMs are also applicable to the investigation of dispersal of invasive species. Václavík et al. (2012) modelled the distribution of the invasive forest pathogen *Phytophthora ramorum* and accounted for different scales of spatial autocorrelation. They assume that accounting for spatial patterns at multiple scales enhanced the understanding of processes that explain ecological mechanisms of invasion while improving predictive performance in static modelling.

Finally, SDM can support decision makers with regard to environmental impact assessment and land use planning or to determine suitable locations for habitat restoration and species reintroduction and thus are a volatile field of research

1.2 MOTIVATION AND RESEARCH QUESTION

Within the project framework of the Alpine Seed Conservation & Research Network the task is to collect seeds from 100 vascular plant species of high conservation value. Some of these species are particularly rare while others are quite common. It often proves difficult to find species populations of sufficient size that yield enough seed material (10,000 seeds per species) although access to the unpublished distribution atlas of the Styrian flora (Niklfeld & Englisch 2004) is available. The atlas is partially incomplete but contributes to the floristic mapping of Central Europe.

The focus is set on six species from our target list, especially Eastern Alps endemics but also some others. This thesis is a feasibility study to assess the opportunities of SDMs to support research in Graz.

Species location data from different localities within the Styrian Alps combined with environmental data will build the basis for the distribution models.

A wide range of statistical and algorithmic tools for species distribution modelling are available today. While all of these methods have their advantages and disadvantages, this study utilizes generalized linear models (GLM) and the machine learning maximum entropy algorithm (MaxEnt) because they are frequently used in modelling species distribution.

The overall concern of this thesis is to identify an optimal approach (GLM, MaxEnt) to predict the distribution of six species scheduled on the project target list with regard to endangerment in the Styrian Alp region. Additionally, this study attempts to generate distribution maps for the selected vascular plant species which will help to refine field surveys.

By developing these distribution models from species occurrence data and digital environmental maps this study aims to answer the following research questions:

- Using two different modelling approaches (GLM and MaxEnt) and existing environmental maps, is it possible to model the recent distribution of six target species and are the results comparable?
- Can the local conditions be represented optimally by the model's parameters?
- How can the use of species distribution models improve success in collecting seed material for the Alpine Seed Conservation & Research Network?
- Is one approach preferable?

2 THEORETICAL BACKGROUND

Analysing species-environmental relationships has always been a central issue in ecology (Guisan & Zimmermann 2000) and most of the modelling approaches developed for predicting species distribution are rooted in quantifying species-environmental relationships (Guisan & Thuiller 2005). Elith & Franklin (2013) define SDMs as quantitative, empirical models of species-environment relationships that are typically developed using species location data (abundance, occurrence) and those environmental variables thought to influence species distributions.

Species distribution models are based on biogeographical and ecological theory and concepts (Franklin 2010) and several decisions (Figure 1) have to be made in advance. The theoretical concept formulates which abiotic and biotic factors are relevant for the species' distribution and on which model assumptions the model is based. Independent sets of species occurrence data are needed to both calibrate and evaluate the distribution model. Environmental variables are usually derived from digital maps and represent the factors that are considered to control species' distribution. The modelling framework addresses different model approaches that link occurrence data to environmental predictors. Finally Data and criteria are needed to evaluate the prediction.

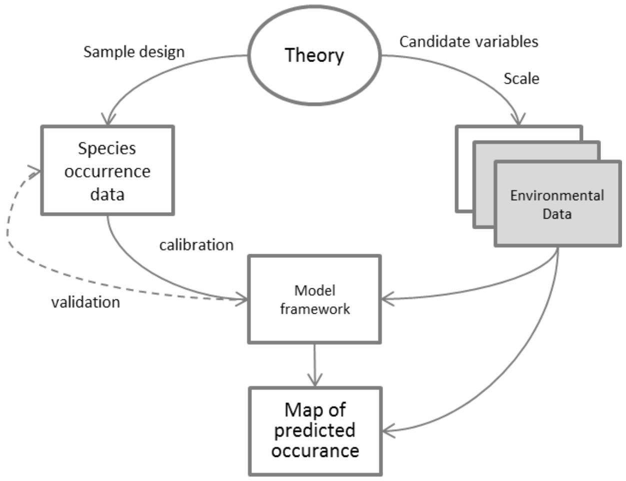


Figure 1: Components of species distribution modelling (Franklin 2010).

Ideally species distribution models are based upon five major steps (Guisan & Zimmermann 2000) that involve (1) conceptual model formulation, (2) statistical model formulation, (3) model calibration, (4) model prediction and (5) model evaluation.

2.1 CONCEPTUAL MODEL FORMULATION

One characteristic of SDM is that they are based on niche concepts (Guisan & Zimmermann 2000; Guisan & Thuiller 2005) or rather niche theory and gradient analysis (Franklin 1995). Concerning the theory and the assumptions behind species distribution models many authors (e.g. Guisan & Thuiller 2005) highlight the importance of the ecological theory that underpins the decisions made at all stages of model development.

In this context, consideration of the main ecological drivers is of particular importance. The causal ecological parameters can either be classified as proximal (direct) or distal (indirect). Direct variables (temperature) directly affects species distribution while indirect variables (elevation, aspect) have no physiological impact on, e.g. plant growth but are linked with causal factors (Austin 2002). Resource gradients are related to matter and energy (water, nutrients, light) consumable by plants and animals (Guisan & Zimmermann 2000). Moser et al. (2005) have shown that energy-driven processes are the primary determinants of vascular plant species richness in temperate mountains. The simplified conceptual model in Figure 2 is adopted from (Franklin 2010) and shows how indirect predictors influence direct and resource gradients which in turn act as main drivers for plant growth. The environmental regime form the fundamental niche (Hutchinson 1957) which results in the realized niche due to biotic interactions (competition) and disturbance (e.g. land use).

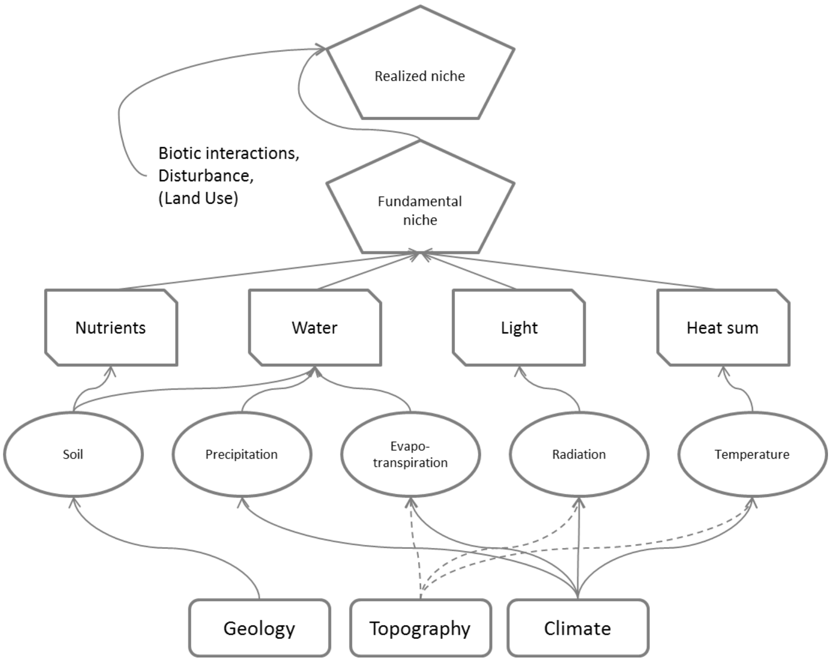


Figure 2: Conceptual model of environmental factors (Franklin 2010).

The assumption that species are in pseudo-equilibrium with their environment is a convenient postulate in species distribution modelling. Static modelling is a valid and powerful approach when species distribution should be modelled with high precision at a large spatial scale under present environmental conditions (Guisan & Zimmermann 2000). These limitations are less restrictive for species, or communities, which are relatively persistent or react slowly to variability in environmental conditions (e.g. arctic and alpine) (Guisan & Zimmermann 2000). However, dynamic modelling approaches were already proposed by (Guisan & Thuiller 2005) and will become increasingly important in the future with regard to global change. Also the simplification, that SDMs quantify Hutchinson's realised niche (Hutchinson 1957) due to the fact that the observed distributions is already constrained by biotic interactions and limiting resources, is common in SDM literature (Guisan & Thuiller 2005).

2.2 STATISTICAL MODEL FORMULATION

Guisan & Zimmermann (2000) stated that the model formulation process addresses two major goals. That is, (1) the choice of a suited algorithm for predicting a particular type of response variable and estimating the model coefficient, and (2) to find an optimal statistical approach with regard to the modelling context.

Today ecologists rely on a diverse range of analytical approaches due to increasing availability of software to implement these methods and a greater computational ability of hardware to run them (Hegel *et al.* 2010).

The methods pursue different approaches and can roughly be divided into three categories (Elith *et al.* 2006; Franklin 2010).

- Regression models, like the generalized linear models (GLM) use relationships between presence or absence and environmental variables. Absences can be simulated using pseudo-absences. They have widely been used in species distribution modelling since the 1980s and the early 1990s (Franklin 2010) and are still common in this field.

- Envelope models, like BIOCLIM or DOMAIN only use presence information and characterize sites that are located within the environmental hyper-space occupied by the species and predict their distribution based on similarity of occurrences.

- Machine learning or statistical learning methods produce rules based on observations and environmental conditions to predict species distribution. MaxEnt compares probability densities from background and presence locations to derive the probability of occurrence.

However, a detailed methodological review is given e.g. in Franklin (1995), Guisan & Zimmermann (2000), Franklin (2010a) or Hegel *et al.* (2010).

This study utilizes generalized linear models as well as the machine learning maximum entropy algorithm to model the geographical distribution of vascular plant species. Thus the next two sections have a closer look to these methods.

2.2.1 GENERALIZED LINEAR MODEL (GLM)

The logistic regression is the most commonly used form of generalised linear model (GLM), and is well suited and widely used in SDM, because it deals with multiple predictors, non-linear response functions and binary response variables (Franklin 2010). GLMs are suitable for distributions such as Gaussian, Poisson, Binomial or Gamma according to the appropriate link function identity, logarithm, logit or inverse (Guisan & Zimmermann 2000). In logistic regression, the dependent variables were subjected to a logit transformation (Hastie *et al.* 2009). Thus the response variable can only take values between 0 and 1. The response variable for the modelling approach is binary (presence/absence) and the response function binomial, thus the logistic regression has been chosen.

Generalized Linear Models are parametric models. The regression models relate a response variable (species occurrence) to a single (simple) or a combination (multiple) of environmental predictor variables (explanatory variables) (Guisan & Zimmermann 2000; Dormann 2012). Link-functions describe the way in which response variable and the explanatory variables are connected (Dormann & Kühn 2009). The GLM can be expresses as:

$$g(y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \varepsilon$$

where y_i is the predicted value at data point i, X_{1i} etc. the values of the explanatory variables at data point i, β the vector to estimate for every model parameter. The link-function $g(\quad)$ describes how the mean of y depends on the linear predictor (Franklin 2010). For binomial distributed data the logit-link is the standard setting and is defined as followed.

$$g(y) = ln\left(\frac{y}{1-y}\right)$$

For minimum/maximum parameter values this equation is approximately 0 or 1. The ideal set of parameters is determined by the Maximum Likelihood (Dormann & Kühn, 2009).

2.2.2 MAXIMUM ENTROPY (MAXENT)

MaxEnt is a machine learning method for making predictions or inferences from incomplete information (Phillips *et al.* 2006) and has been developed in the field of statistical mechanics. The algorithm is suitable for presence only (PO) data and can deal with problems of small samples that have not been designed (sample bias) (Phillips *et al.* 2006; Franklin 2010).

The principle of maximum-entropy states, that a probability distribution, subject to known constraints is the best approximation of an unknown distribution because it agrees with everything that is known (close to uniform) but avoids everything that is not known (Phillips *et al.* 2006; Franklin 2010). The unconstrained distribution is that of all factors in the study area, and the constraint is that the expected value is approximated by an empirical set of observations of species presence.

From a statistical viewpoint, MaxEnt minimizes the relative entropy between two probability densities (estimated from presence data and from landscape) defined in feature (covariate) space (Elith *et al.* 2011). That is, that the algorithm extracts background data from landscape and contrasts them against the presence locations.

$$\Pr(y = 1|z) = f_1(z)\Pr(y = 1)\,/f(z)$$

The equation above shows that, if the conditional density of the covariates is known at the presence sites, $f_1(z)$, and the density of covariates across the study area $f(z)$ are also known, then only the prevalence Pr(y=1) is needed to calculate the conditional probability of occurrence (Elith *et al.* 2011).

The logistic output format in MaxEnt is a post-transformation of MaxEnt's raw output (Elith *et al.* 2011) and gives the probability of occurrence (Phillips & Dudík 2008) which is the easiest to conceptualize format. The logistic output format gives values between 0 and 1 of probability of presence. Typical presence locations have a probability of presence of about 0.5 (Phillips 2008).

MaxEnt deals with sample selection bias by using target-group background (Phillips & Dudík 2008). The uniform background data is replaced by a random sample of the biased sampling distribution. The result is that both background and species presence is biased in the same manner and that MaxEnt has much better predictive performance.

Although MaxEnt is similar to GLM (Phillips *et al.* 2006), an important distinction between MaxEnt and logistic regression models is that MaxEnt does not interpret locations without species occurrences records as absence, but rather as representing the background environment (Franklin 2010). This means that background locations are not interpreted as absences.

The MaxEnt modelling framework is very functional because it offers several thresholds, statistics for model evaluation and is able to compute the importance of specific environmental variables.

2.3 MODEL CALIBRATION

The process of deciding on and selecting the explanatory variables or combination of variables that should be included in the model is called model calibration (Guisan & Zimmermann 2000). The choice of appropriate environmental predictor variables is a crucial step in species distribution modelling to ensure accuracy and model realism (Mod *et al.* 2016). The challenge is to identify the environmental predictors (usually derived from digital maps) that represent the resource gradient or other factors that determine the species distribution at an appropriate scale (Franklin 2010). Ecological parameters that are believed to be the causal driving forces for the distribution (and abundance) of a species (Guisan & Zimmermann 2000) have to be identified. Neglecting ecological knowledge or ecophysiologically meaningful predictors limit the predictive power of statistical species distribution models and results in incomplete niche quantification (Austin 2002; Mod *et al.* 2016). On the other hand, reducing the number of predictors to a reasonable quantity enhances the accuracy and the predictive power of a model (Guisan & Zimmermann 2000), therefore available environmental variables need to be evaluated with regard to ecological realism.

The variable selection can either be done arbitrarily or automatically using stepwise selection (Guisan & Zimmermann 2000) or other approaches suggested e.g. by Guisan & Thuiller (2005) like multi-model inference, boosting and model averaging, shrinking methods or hierarchical partitioning.

In multiple regressions, like GLM, multicollinearity can weaken significance or reverse the sign of proximal variables due to the fact that a correlated variable is already present in the model (Franklin 2010). Thus, multicollinearity is omnipresent throughout the entire model selection process. Strongly correlated variables usually are excluded with preference to the ecological meaningful one.

The fitting of a model usually goes hand in hand with a reduction of variance (Guisan & Zimmermann 2000) or deviance in GLMs. For GLMs the adjusted D^2 or R^2 is an ideal measure for model comparison (Guisan & Zimmermann 2000).

The Akaike Information Criterion (AIC) is also frequently used in the model selection process (Franklin 2010). The criterion is a measure of goodness of fit and accounts for the number of parameters. A smaller AIC (lower unexplained deviance) indicates a better model.

2.3.1 Collinearity and Spatial Autocorrelation

Environmental factors often act in similar ways that is that they are not independent of each other (Dormann *et al.* 2013). This causes two major problems (Dormann 2012). Firstly, the important variable could not be determined because of similar acting variables overlap. Secondly, collinear predictors lead to unstable estimates. Unfortunately there is no specification given in statistic literature describing a definition of “highly correlated” (Dormann 2012). A threshold of 0.7 is commonly used but has also been chosen more (0.4) and less restrictive (0.85) (Dormann *et al.* 2013). In fact, this problem can hardly be removed, but the higher the correlation is, the harder is the parameter estimation and the standard error increases. This effect is called variance inflation and could be quantified via the variance inflation factor (VIF). A VIF score greater than 10, indicates problems in the regression (Dormann 2012). VIF is a valid diagnostic for GLM distribution models (Guisan & Zimmermann 2000).

Because species distribution models are dealing with spatial data they are complicated by spatial autocorrelation (SAC). Spatial autocorrelation means locations close to each other exhibit more similar values than those further apart and this refers directly to what Tobler (1970) calls “The First Law of Geography”. One of the key assumptions of standard statistical analyses, namely, that the model residuals are independent and identically distributed, is violated if SAC remains in the residuals of a species distribution model (Dormann *et al.* 2007). This may bias parameter estimates and can increase type I error rates (false rejecting the null hypothesis of no effect). Even if it is easy to prove spatial autocorrelation, its elimination is complex especially in a predictive setting (Dormann *et al.* 2007).

One possible opportunity is to exclude observations within a certain minimum distance as Parolo et al. (2008) did to minimize SAC through the use of a constrained random split of sampling data.

However, Dormann et al. (2007) reviewed six statistical methods for different types of distribution that account for spatial autocorrelation in model residuals. Assuming the state of spatial stationarity and isotropic spatial autocorrelation, they found out, that most of these spatial modelling techniques showed good type I error control and precise parameter estimation using simulated data with known autocorrelation in the model residuals. Generalized linear mixed models (GLMM) together with spatial generalized estimating equation (GEE) and spatial eigenvector mapping (SVEM) are the most flexible methods used for addressing SAC for various error distributions (Dormann *et al.* 2007). Geographically Weighted Regression (GWR) is another geostatistical approach that accounts for spatial autocorrelation. GWR has limited use for hypothesis testing (Dormann *et al.* 2007; Franklin 2010) and is not designed for removing SAC (Dormann *et al.* 2007). Spatial eigenvector mapping and the autocovariate (AC) method are examples of two of the methods that are suitable for both MaxEnt and GLM because they can easily be added as an additional covariate into the model. However, the problem of model prediction is still complex.

2.4 DATA USED FOR MODELLING SPECIES DISTRIBUTION

Model development involves both biological data and environmental data. In SDMs species observational data forms the response (dependent variable). For species this can be counts, cover-abundance estimates, presence-absence or presence-only records (Elith & Franklin 2013). If it is possible to undertake extensive field surveys one can identify real absences as well. However, in many cases one is confronted with presence only data, for example when the data comes from atlases, databases or herbarium material. If that is the case pseudo-absences are needed to simulate species' absence. These are randomly generated points within the study area that are considered to be absence but have not actually been visited. But there is some uncertainty because pseudo-absences can be located in regions where the species is present.

Species distribution models use variables that are related to the primary environmental regimes of heat, light, moisture and nutrients (Franklin 2010). Microclimatic features seem to be the key-factors for the understanding of the relation between alpine species and global warming (Scherrer & Körner 2011; Patsiou *et al.* 2014; Maclean *et al.* 2015). Thus species distribution models optimally include both broad scale climate variables and fine scale terrain information, which results in local variation in water, energy and nutrient availability also being added to the model (Franklin 2010).

Environmental data is typically gathered from digital maps. In many cases, digital environmental maps are derived from interpolations, calculations or combinations and thus are less precise than the maps from which they originate (Guisan & Zimmermann 2000). Digital elevation models (DEM) and its derivatives are among the most accurate environmental predictors, but not always in ecological sense.

Another essential aspect concerning species data is sampling design. The number and geographic distribution of samples is important in order to model the response to environment in its complexity (Elith & Franklin 2013). Sampling design might be of particular relevance to cover the whole ecological gradients in the study area (Guisan & Zimmermann 2000). It ensures a set of unbiased and representative data. Because of several reasons sample data is not independent and unevenly distributed (Elith & Franklin 2013). For example presence points are more often located closer to hiking paths.

2.5 MODEL PREDICTION

Since the ecological profile was modelled with any of the available techniques it is possible to predict the species' potential distribution (Guisan & Zimmermann 2000) for GIS-implementation.

2.6 MODEL EVALUATION – AUC

The model validation verifies whether the model meets certain criteria and whether it is acceptable for its purposes. A number of methods and criteria have been developed.

The threshold-independent area under the receiver operating characteristic (ROC) curve, known as AUC, is a standard method used to assess the accuracy of predictive distribution models. It offers an appropriate measure to quantify model performance in particular when comparing different models, predictor combinations or species.

In a ROC plot (Figure 3), the true positive proportion (sensitivity) is plotted against the false positive proportion (1 - specificity) (Fielding & Bell 1997). Sensitivity and specificity are measures of classification accuracy derived from a confusion matrix.

The AUC (area under the curve) is calculated by adding the area under the ROC curve and is often considered an important index because it provides a single measure for overall accuracy (Fielding & Bell 1997).

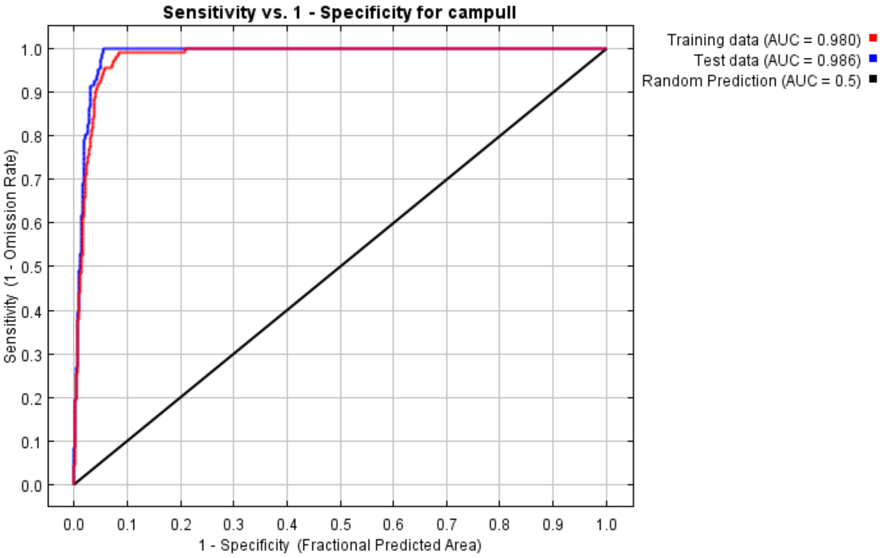


Figure 3: ROC curve for Campanula pulla.

The AUC has been criticized i.a. because it is dependent on the geographical extent of the study area and does not account for the spatial error distribution (Lobo *et al.* 2007).

However, the AUC score is still an appropriate measure for overall accuracy in the model evaluation process.

2.7 THRESHOLD METRICS

Thresholds are needed to convert continuous prediction maps into binary maps that are used in many practical applications (Lobo *et al.* 2007; Franklin 2010). However, threshold dependent metrics are often considered to be subjective (arbitrary).

Lobo *et al.* (2007) suggested a threshold which sensitivity and specificity are equal. The rate of true positives decreases while the rate of true negatives increases when the threshold changes from 0 to 1. The crossing point of both curves is the optimal threshold. This intersection is optimally located in the north-western most point of the ROC curve (Figure 4). In MaxEnt this threshold is called “Equal training sensitivity and specificity”.

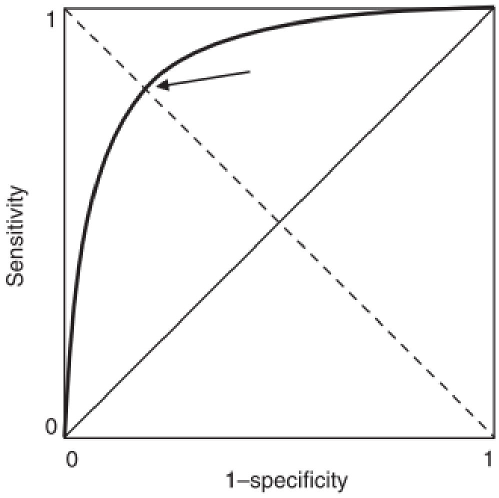


Figure 4: ROC curve showing the most north-western point (arrow), where specificity=sensitivity (Lobo *et al.* 2007).

3 MATERIAL & METHODS

This study uses the framework of the above mentioned theoretical background and utilizes static, empirical distribution models that assume pseudo-equilibrium without taking competition and interactions into account. Two different modelling approaches (GLM and MaxEnt) were utilized to derive predictive distribution maps of six vascular plant species. The models were calibrated with a set of topo-climatic predictor variables and stepwise model selection for GLM models using AIC scores.

MaxEnt used a jackknife test to specify variable importance.

For GIS implementation probabilistic prediction maps were created and evaluated using AUC-score. Based on the threshold where sensitivity and specificity is equal, binary maps were produced. To quantify the results the area of occupancy was calculated for each species.

This section details the methodological procedure and the raw data.

3.1 STUDY AREA

This thesis refers to the distribution of alpine vascular plant species in Styria. The study therefore is restricted to the elevated regions of Styria (Eastern Alps). The physiogeographic and climatic (Harlfinger 2010) conditions within the Styrian Alps is heterogeneous and thus only a brief overview is provided to highlight the important parameters for plant growth.

The Alps stretch over an arc of more than 1,200 km from the Mediterranean cost to the Vienna basin covering an estimated area of 200,000 km^2 (Ozenda & Borel 2003). Styria is located in the eastern region of the Alps. The highest summit is Hoher Dachstein (2995 m) located at the border to Upper Austria.

Geological and climatic pattern are fundamental determinants of species distribution, including the alpine belt (Ozenda & Borel 2003). Geological bedrock and climatic conditions show remarkable differences which results in a specific flora between the central and the limestone Alps.

Higher altitudes in the northern Alps are directly affected by northern and western weather events. This causes a mountain climate with high precipitation and snow rich winters. Contrastingly, the northern slope of the Niederen Tauern is shielded by the northern Alps and the conditions are slightly weakened in terms of levels of precipitation (summer and winter), but not in terms of frequency. The main ridge of the Styrian central Alps often act like a climatic divide between the northern and the southern side. Higher altitudes of the central Alps display a strong central alpine climate with relative low precipitation and snowfall. The eastern edge of the Styrian Alps is strongly affected by south and south-eastern weather events. This results in a higher rate of thunder storms with hail.

Hypsometric temperature reduction is reduced in the central alpine part of Styria due to greater mountain massifs. Thus the altitudinal zones (tree line, snow line) are shifted upwards. However, most Alp regions (including Styria) this natural zonation is absent due to anthropogenic impact. The tree line has been depressed by estimated 200–300 m and closed alpine grasslands (the alpine altitudinal belt covers nearly 15,000 km^2 within the whole Alps) are best developed in the Alps (Ozenda & Borel 2003).

3.2 MODEL INPUT DATA

Species data

This study uses existing field data from the Austrian Vegetation Database. The aim of this database is to gather phytosociological information (vegetation relevés) from the territory of Austria (Willner *et al.* 2012). The database contains a large part of entries between the years 1997 and 1999 and between 2001 and 2003 and is still relevant for current studies.

Species data is managed in the non-commercial software package JUICE which is widely used for editing and analysing of phytosociological data (Tichý 2002).

Unfortunately, the data collected within this database is fairly heterogeneous in regard to plot size and completeness. Therefore, only the geographic location has been selected for the species distribution models. Each relevé that contains the species of interest is supposed to be a presence location, regardless of plot size or date of the relevé. Unfortunately, a certain positional error must also be accounted for because this has not always been documented.

Sample size must be large enough to make meaningful statistical inferences. While MaxEnt is suitable for low sample sizes GLM needs a higher number of observations.

As a primer criterion for species selection, the frequency within the database has been chosen. Only a sufficient number of presence locations yield good model results. A second criterion held certain level of relevance in nature conservation. The following species have been modelled for this study:

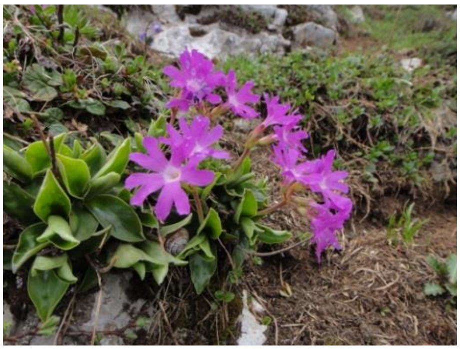


Figure 5: Primula clusiana

Code: priclus

Primula clusiana is an endemic vascular plant species of the north eastern limestone Alps and could be found from the montane to the alpine zone. The calcicole species could rarely be found in the central Alps. Preferred habitats are moist and rocky grassland and snow bed communities. Photo: C.Berg

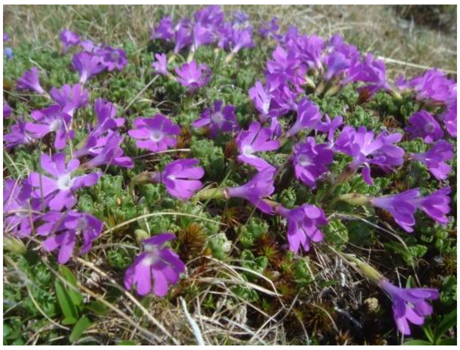


Figure 6: Primula minima

Code: primini

Primula minima is a species from siliceous neglected grassland (Caricetum curvulae) in the subalpine to alpine zone and can also be found in snow bed communities. The calcifuge species is rather common in the central alps but could also be found on low-lime humus soils in the northern limestone Alps. Photo: C.Berg

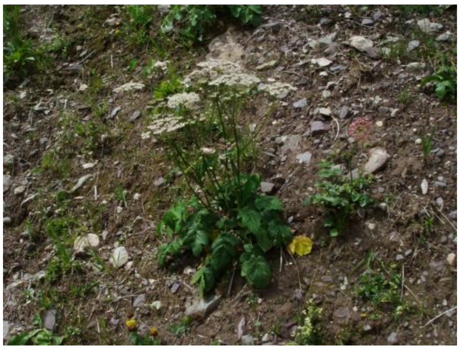


Figure 7: Heracleum austriacum

Code: heraust

Heracleum austriacum is frequent from the subalpine zones of the limestone Alps. It is an eastern Alps endemic species that grows in tall forb communities, scree, stony subalpine meadows and krummholz. Photo: C.Berg

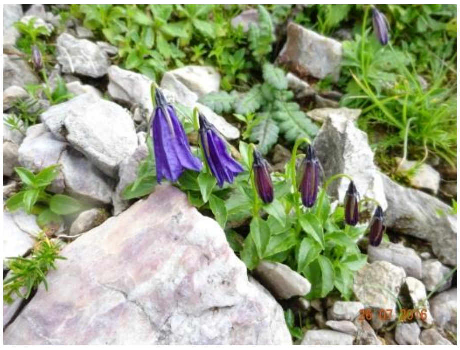


Figure 8: Campanula pulla

Code: campull

Campanula pulla is a calcicole eastern Alps endemic plant species. It is common in the northern limestone Alps but is rare in the central Alps. The altitudinal distribution ranges from the subalpine to the alpine zone. It grows in moist scree habitats and snowbed communities. Photo: P. Schwager

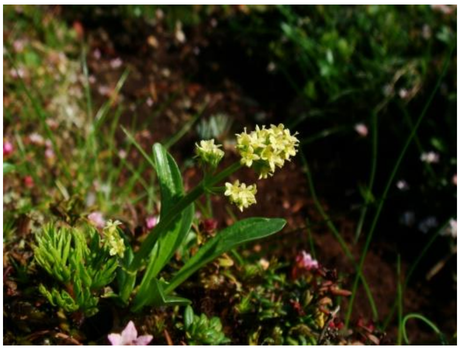


Figure 9: Valeriana celtica

Code: valcelt

As a calcifuge species Valeriana celtica is mainly distributed within the subalpine and alpine zones of the central Alps but extends its range of occurrence to the northern limestone Alps where it present but sparse. Photo: C.Berg

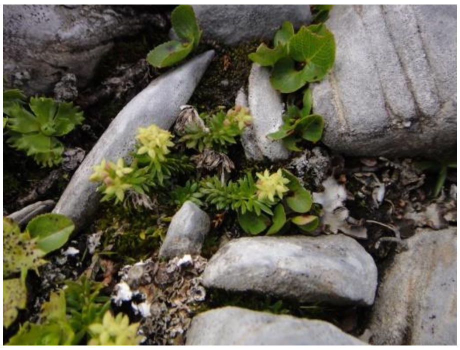


Figure 10: Galium noricum

Code: galnori

Galium noricum is a common, sub endemic species in the north eastern limestone Alps. From the subalpine to alpine zones the calcicole species grows in rocky meadows, scree and could often be found near the summit regions. Photo: C.Berg

Floristic and location specific details as well as information of endemic status is taken form Maurer *et al.* (1989), Maurer (1996), Maurer (2006), Aeschimann *et al.* (2005) and Fischer *et al.* (2008).

Environmental Data

Environmental data used in this study is freely available at the Austrian open data portal (www.data.gv.at).

The “Klimaatlas Steiermark” (Harlfinger 2010) provides climate data for the period 1971 to 2000 which is available in a resolution of 50 meters for the whole Styrian area in Asci format.

Geological data has been gathered from a shape file (1:200 000) that was originally provided from the Geologischen Bundesanstalt Austria.

The available digital elevation model (DEM) in Asci format has a resolution of 10 m and was resampled to a coarser resolution of 50 m.

A preselection of the entire set of environmental predictor variables (more than 100) was conducted in order to exclude variables without ecological relevance and variables that were highly correlated. Table 1 show the variables used for the models.

There is some evidence of collinearity between average annual temperature (t_jahr_k) and average summer precipitation (rrsum_somm) (Figure 11). However, both variables are important for plant growth and were used for the models. During the model calibration process in section 3.4.3, the variance inflation factor (VIF) for each final GLM model was calculated using the vif() function in the “car” package (John *et al.* 2015).

Table 1: The environmental data used for the models.

Name	Description	Range [unit]	origin
slope	Slope	0,004 – 0,729 [rad]	DEM www.data.gv.at
northing	cos(aspect)	-1 – +1	DEM www.data.gv.at
easting	sin(aspect)	-1 – +1	DEM www.data.gv.at

gesteine	Geology, classified into calcareous or not calcareous	0 / 1	ESRI Shapefile www.data.gv.at
rrsum_somm	Average sum of precipitation in summer	318 – 598 [mm/Monat]	Klimaatlas Steiermark www.data.gv.at Interpolation
globre_jr	Average annual global radiation at real surface.	563 – 1310 [kWh/m^2]	Klimaatlas Steiermark www.data.gv.at Interpolation
t_jahr_k	Average annual temperature.	274.2 to 282.34 [K]	Klimaatlas Steiermark www.data.gv.at Interpolation

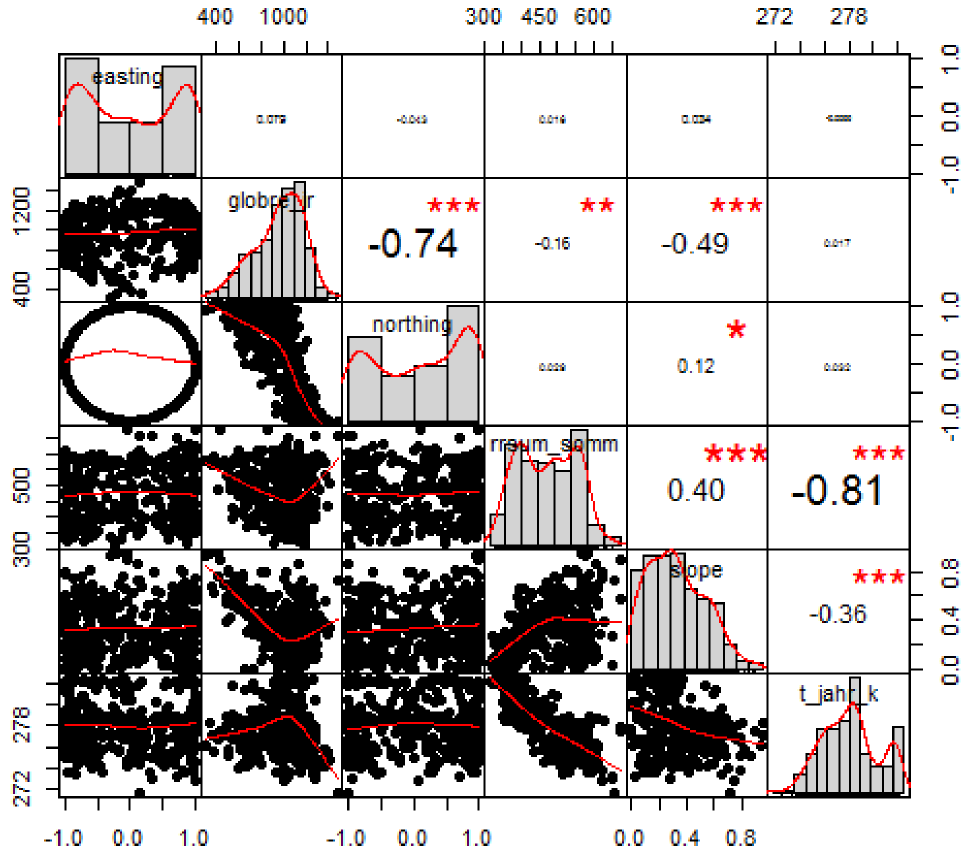


*Figure 11: Correlation matrix. Significance codes 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Description of the variables used:

- Slope (slope): The steepness of slope is related to overland and subsurface flow of water and affects soil characteristics like moisture, texture or development (Franklin 2010). Slope was calculated from the DEM using the terrain analysis tool “Slope, aspect, curvature” (SAGA 2.3.2) implemented in QGIS. The algorithm uses a “9 parameter second order polynom” and is described at Zevenbergen & Thorne (1987).

- Aspect (northing, easting): The direction the slope is facing, affects the amount of solar radiation received on the slope and the seasonal and annual patterns of solar insolation (Franklin 2010). This in turn affects soil moisture availability. The variable is an indirect proxy for evapotranspiration because on southern facing slopes there is greater insolation. Aspect was calculated simultaneous with the SAGA tool “Slope, aspect and curvature”. For the statistical analysis “aspect” was converted to the circular variables “northing” and “easting”. “Northing” is calculated as cos(aspect) and “easting” as sin(aspect).

- Geology (gesteine): Geology (and soil) are indirect proxies for nutrient and water availability as well as for chemical properties of the substrate (e.g. pH). To integrate the factor geology in the models the detailed geological entities were assigned to the classes calcareous and non-calcareous. The geological data is categorical information and is optimally be aggregated into the smallest number of classes that are ecologically relevant (Franklin 2010). Finally the vector data set has been rasterized.

- Average sum of precipitation in summer (rrsum_somm): Although precipitation is a relatively poor surrogate for plant water availability (Mod *et al.* 2016) average annual or some seasonally average have been used in most of the SDM studies (Franklin 2010).

- Average annual global radiation at real surface (globre_jr): Global radiation is the sum of diffuse and direct radiation. Solar radiation represents heat rather than photosynthetically active radiation and acts similar than temperature (Mod *et al.* 2016).

- Average annual temperature (t_jahr_k): Temperature is one of the most frequently used predictors in SDM research and is of particular importance for plant growth. The average annual temperature, however, does not represent the growing season or over-winter conditions, which are essential in plat distribution (Mod *et al.* 2016). Seasonal averages (e.g. winter temperature) are complicated by multicollinearity.

3.3 SOFTWARE

Data manipulation and management was primarily completed using QGIS (version 2.18.7) and R (version 3.3.1) in combination with R-Studios (version 1-0-136). The statistical modelling made use of R and MaxEnt.

QGIS

QGIS is a user friendly Open Source Geographic Information System (GIS) licensed under the GNU General Public License and is an official project of the Open Source Geospatial Foundation (OSGeo). It is platform independent and can be run on Linux, Unix, Mac OSX Windows and Android. It supports a variety of vector, raster and database formats and the functionality can be extended by a variety of toolboxes (e.g. GRASS GIS, SAGA GIS, R).

QGIS is freely available at http://www.qgis.org

MaxEnt

MaxEnt is an open source stand-alone Java application for geographic species distribution modelling. MaxEnt can also be used in combination with R. It is implemented through the R packages “maxnet” and is also available in the “dismo” package.

It is freely available from the Website of the American Museum of Natural History http://biodiversityinformatics.amnh.org/open_source/maxent

R-Project

R is the open source version of S, a script language and environment for statistical computing and graphics.

R provides a large library for statistical techniques like linear- and nonlinear modelling and classical statistical tests and is highly extensible.

R runs on a variety of platforms and is freely available at https://www.r-project.org/

3.4 WORKFLOW

The Workflow can be separated into four major steps (Figure 12). In a first step the raw data was pre-processed to get a consistent set of species and environmental data. In the second step the training and testing data set was generated. The third step encompasses the creation of the model as such and the fourth step includes creating a prediction that generates the prediction maps. Finally the distribution map was evaluated.

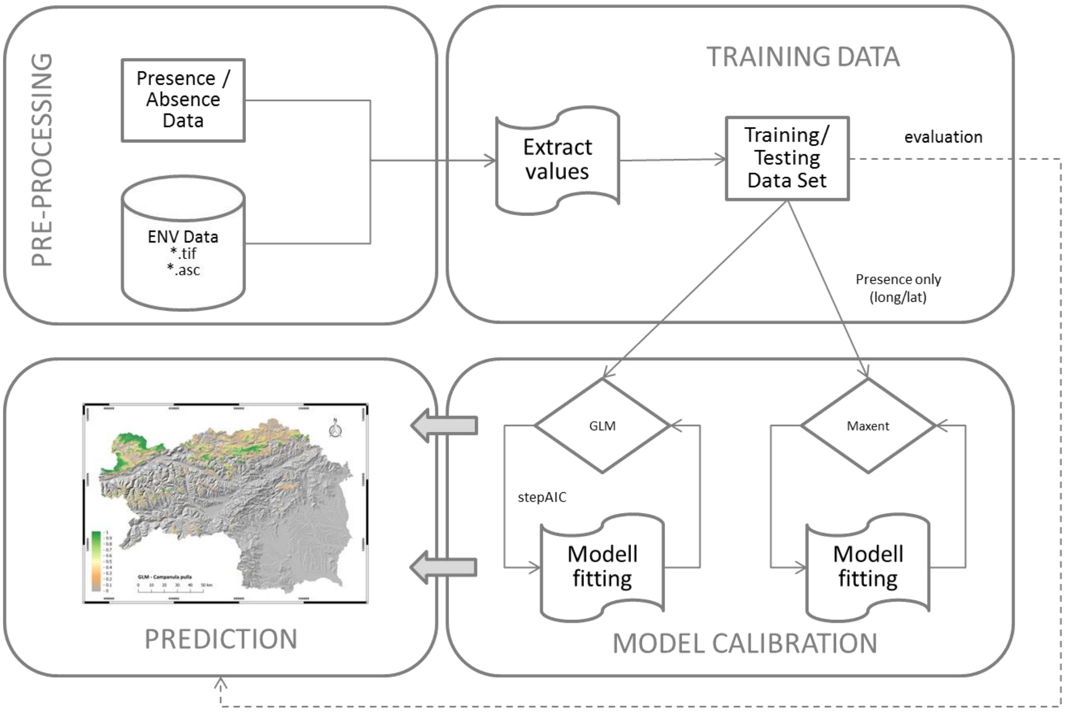


Figure 12: Workflow for model building.

3.4.1 Pre-processing

Presence / Absence data

The species data was stored in a Juice table and has been exported into the data exchange format comma separated values (*.csv). This file contains all site information that was noted in the field (e.g. coordinates, or cover of different plant layers).

For this study only the coordinates of each presence location were used and thus the raw data set was cleaned in Excel.

The presence points are mostly clustered within some parts of the study area, thus a grid of 500 m resolution was used to randomly select one point per grid cell. Moreover independent training and testing data sets are necessary for model evaluation so the remaining presence points were used to evaluate the models.

The generalized linear models (GLM) need additional pseudo-absence information. Pseudo-absence points (500 points) were randomly generated within the whole Styrian area. These pseudo-absence points were then subdivided by grouping them randomly to either testing or training absence points.

Finally, three different data sets are necessary for training and evaluating the GLM model (Figure 13). The first data set contains both presence and pseudo-absence points and was used to train the model. Presences were coded with 1, absences as 0.

Another data set contains the remaining presence points that were not selected with the 500 m grid. The last data set contains the pseudo-absence points that were not used for the training data set.

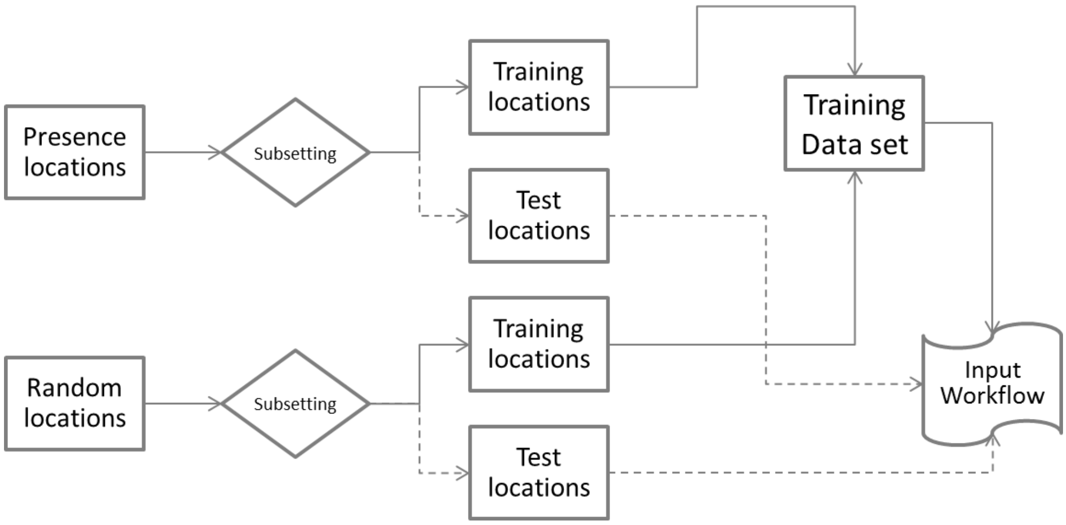


Figure 13: Pre-processing of presence/absence data.

Environmental data

In R raster layers can be organized in a raster Stack. A raster stack is a collection of “RasterLayer” objects with the same spatial extent and resolution. To build a raster stack of predictor variables it was necessary to convert all environmental raster layers to same spatial resolution, extent and coordinate reference system.

This step was automated with a R-script by using functions of the “raster” package (Cheng *et al.* 2016), the “sp” package (Hijmans *et al.* 2016) and the “rgdal” package (Pebesma *et al.* 2017).

The script iterates the folder that contains the environmental raster layers. It initially uses a digital elevation model as a template, then crops and resamples all other raster layers to the same extent and resolution. Finally the new environmental raster layers are saved to an output folder either as Tiff-File (GLM-input) or ASCII-File (MaxEnt-input).

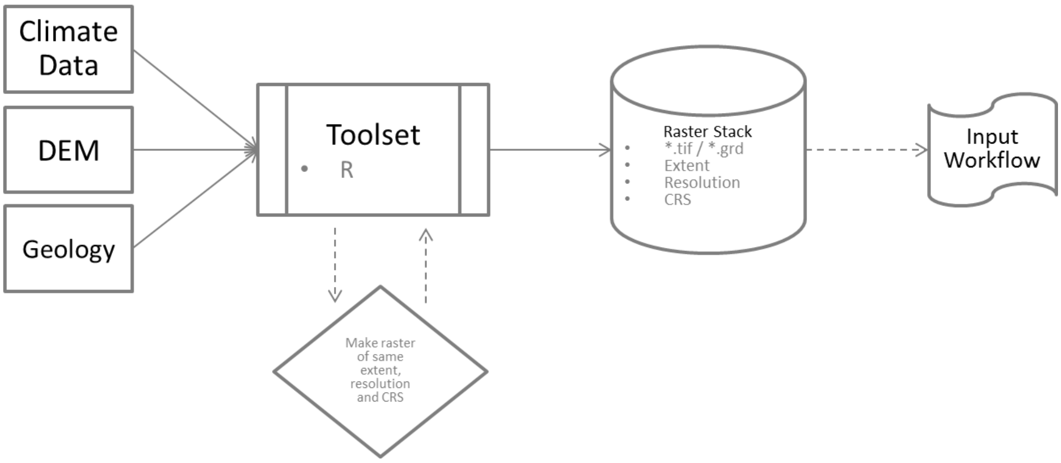


Figure 14: Pre-processing of environmental raster layers to build a raster stack.

3.4.2 DATA EXTRACTION FOR TRAINING THE MODELS

For each point of the three data sets (generated as described in section 3.4.1) the values of the environmental raster layers were extracted to a data frame. Duplicate records and records with NAs were preliminarily excluded. .

The training dataset is subsequently used to train the GLM-model. MaxEnt can use presence-only data so only the presence training locations are needed.

The test and the test-pseudo-absence data set are used to evaluate model accuracy (section 2.6).

3.4.3 MODEL CALIBRATION – GLM

The pre-processing of the data is described in section 3.4.1.

The GLM models were fitted with the pre-processed training data set in R using the R package “dismo” (Hijmans *et al.* 2017). First all potential useful predictors (full model) were included to model each of the six species. The predictors were linked in the models with the following code.

```
model.full<- glm(formula = pb_train ~ gesteine + globre_jr + northing + rrsum_somm + slope +
t_jahr_k, family = binomial(link = "logit"), data = env)

step.model<- stepAIC(model.full, direction=c("both"))
```

A stepwise approach based on Akaike Information Criterion (AIC) was chosen to find the final model with the lowest AIC-score. The “Mass” package (Ripley 2017) offers the stepAIC() function that systematically drops on predictor variable and calculates a model with the remaining variables. The stepwise selection was done in both directions.

The model with the lowest AIC value (low unexplained deviance) is considered to be the best model. The process also includes an “anova” component corresponding to the steps taken by the algorithm. Appendix 7.1 shows the output of the stepwise model selection and Table 4 shows the variables (and significances) that were chosen from the stepwise selection for each final model.

The final models were taken as they were generated by the stepAIC() function. That is, no further reduction of possible insignificant variables was done. Thus the selection remains comprehensible.

The GLM models were tested on collinearity using the VIF-scores. Table 2 details the scores for each variable in the final model that was used for prediction. All predictor variables have low VIF values (Table 2), which indicate that collinearity between the predictor variables is low.

Table 2: VIF scores calculated for used variables in the each GLM model.

Art	gesteine	globre_jr	rrsum_somm	slope	northing	easting	t_jahr_k
campull	1.475	8.548	1.236	2.284	5.876		2.284
galnori	1.343	1.215		1.115			1.496
primini	1.951		2.921	1.142			2.689
priclus	1.255		1.183				1.281
heraust	1.608		1.388	1.095			1.599
valcelt	2.933		3.559	1.136			2.065

3.4.4 MODEL CALIBRATION – MAXENT

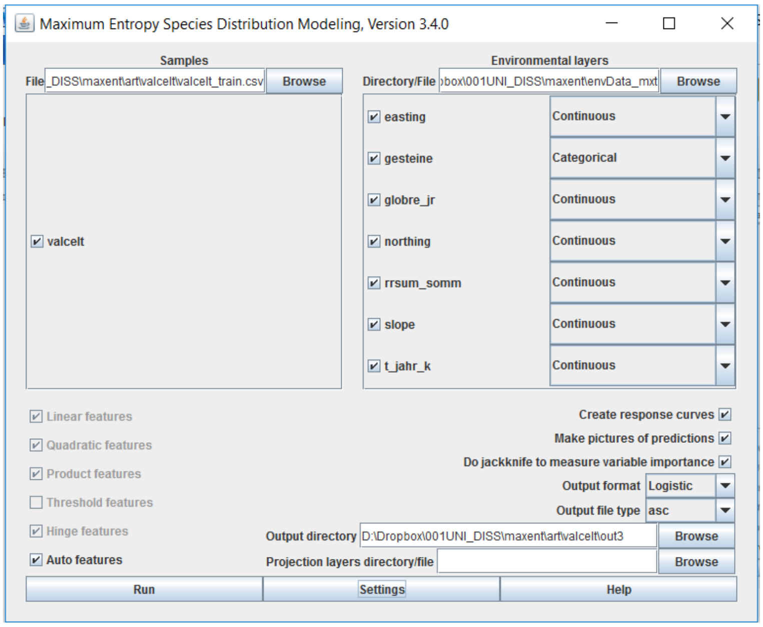
MaxEnt offers a graphical user interface where specific settings can be selected (Figure 15). To create a model in MaxEnt the path to training and testing data needs to be set. Training and testing localities comply with the presence and testing points that were used for the GLM models.

The logistic output format gives the probability of occurrence. To obtain this output, the format was set to “Logistic” in MaxEnts user interface.

The directory for the environmental predictors is specified at “Environmental layers”. It is possible to set the type of variable for each variable. All variables are continuous, except the geological predictor (gesteine), which is a categorical variable.

Finally, an output directory is needed where the output data is to be stored.

The default settings have been applied for all other settings. Reduction of variables was not performed.



Maximum Entropy Species Distribution Modeling, Version 3.4.0

Samples

File _DISS\maxent\art\valcelt\valcelt_train.csv

☑ valcelt

Environmental layers

Directory/File ›box\001UNI_DISS\maxent\envData_mxt

☑ easting	Continuous
☑ gesteine	Categorical
☑ globre_jr	Continuous
☑ northing	Continuous
☑ rrsum_somm	Continuous
☑ slope	Continuous
☑ t_jahr_k	Continuous

☑ Linear features

☑ Quadratic features

☑ Product features

□ Threshold features

☑ Hinge features

☑ Auto features

Create response curves ☑

Make pictures of predictions ☑

Do jackknife to measure variable importance ☑

Output format Logistic

Output file type asc

Output directory D:\Dropbox\001UNI_DISS\maxent\art\valcelt\out3

Projection layers directory/file

Figure 15: MaxEnt graphical user interface.

MaxEnt tracks the variable contribution while training the model (Table 3). The percent contribution (“c” in Table 3, respective Figure 16) of a variable indicates the degree to which the variable affects the model. But the values should be interpreted with caution because they depend on the path that the algorithm takes to find the optimal solution. Additionally, the permutation importance (“i” in Table 3) indicate which variable largely decreases AUC.

A subsequent jackknife test was performed to determine variable importance. The test indicates both which variable has the most useful information by itself and which variable has the most information that is not present in other variables (Phillips 2008).

The process is comparable to a stepwise model selection and is based on “gain” in MaxEnt. Gain is closely related to deviance which is a measure of goodness of fit in GLMs (Phillips 2008).

A number of models were generated using the jackknife test. Each variable is excluded in turn and a model is created with the remaining variables. Then models were created using each predictor in isolation. Finally, a model is created using all variables. The result of the jackknife test is shown in three bar charts. Figure 17 shows the variable importance obtained from the jackknife test in MaxEnt with regard to AUC.

Table 3: Relative contribution of the environmental variables to the MaxEnt model. C: % contribution; I: permutation importance.

Species	**gesteine**		**globre_jr**		**rrsum_somm**		**slope**		**northing**		**easting**		**t_jahr_k**	
	c	**i**	**c**	**i**	**c**	**i**	**c**	**i**	**c**	**i**	**c**	**i**	**c**	**i**
campull	47.4	21.7	0.5	2.7	11.9	12.8	1.5	2.4	0.7	1	0.6	0.5	37.3	58.9
galnori	35.4	13	0.2	0.7	7.5	11	2.1	1.4	0.6	0.5	0.4	0.3	53.8	73
primini	0.1	0.1	0.1	0.1	0.4	0.2	0.6	0.2	0.2	0.1	0.2	0.1	98.4	99.1
priclus	47	31.7	0.1	0.7	9.1	27.3	0.3	0.5	0.4	0.6	0.6	0.2	42.5	36.1
heraust	48.1	19.7	0.2	0.2	21.3	25.6	0.6	1	0.9	0.4	0.4	0.2	28.6	52.7
valcelt	0.4	0.1	0.1	0.1	0.7	0.4	1.7	0.8	0.3	0.1	0.3	0.2	96.5	98.2

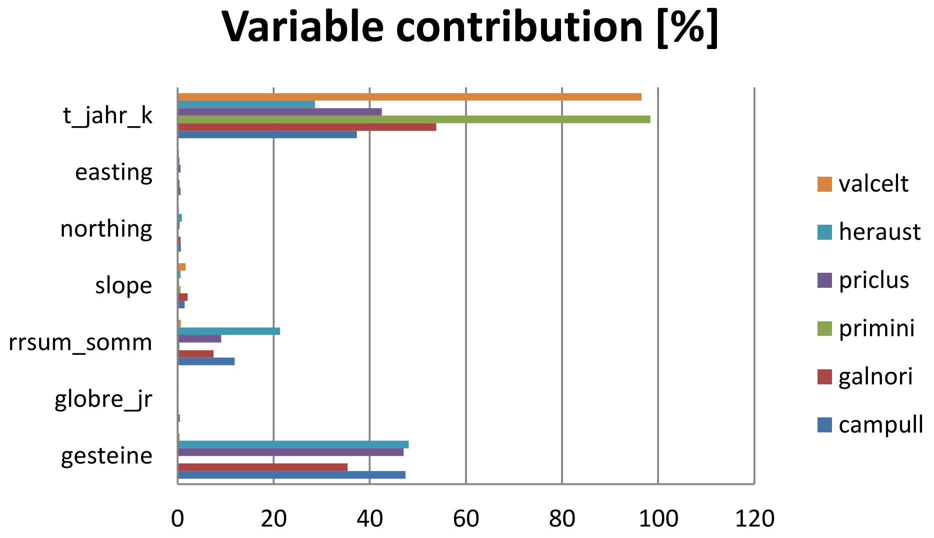


Figure 16: Variable contribution [%] in the MaxEnt models for each species.

*Table 4: Variable significance of the final GLM model for each species. Significance: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

	campull		galnori			heraust			priclus		primini				valcelt		
Estimate	*	***	*	***	" "	*	**	***	**	***	*	***	.	" "	*	***	" "
(Intercept)		5.074		7.271		1.984				3.342		6.589				6.224	
gesteine1		5.259		6.565				6.325		6.992			1.75		2.044		
globre_jr		-3.322			-1.441												
northing	-2.105																
rrsum_somm	1.971						3.081		2.982		-2.388				-2.146		
slope		-4.069	-2.414			1.976								-1.42			-1.495
t_jahr_k		-5.239		-7.307		-2.284				-3.618		-6.73				-6.311	

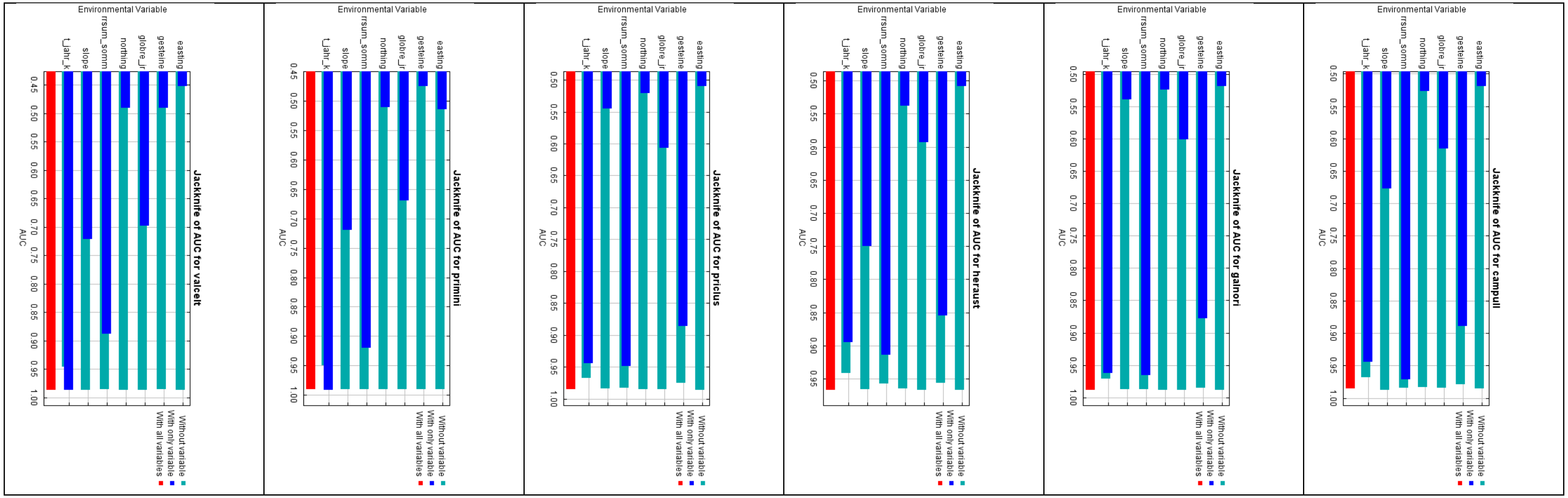


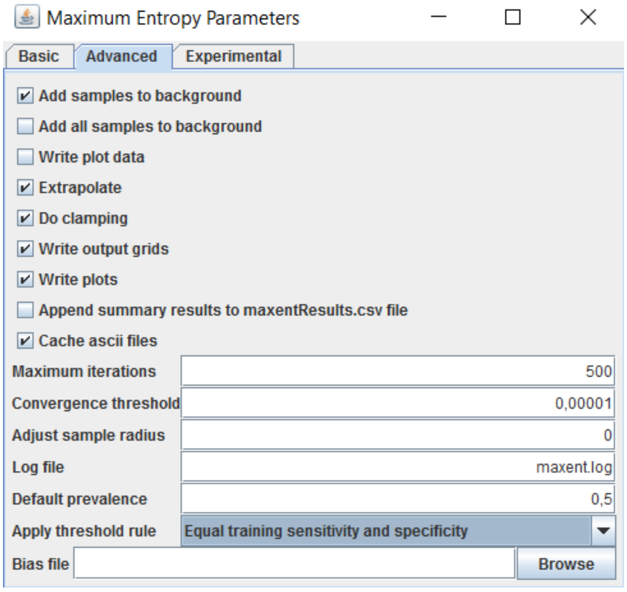
Figure 17: Barplots of the jackknife test on AUC.

3.4.5 Model Evaluation

For both, GLM and MaxEnt the same presence data set was used for evaluation. The pre-processing of the test data set is described in section 3.4.1.

For the GLMs the model evaluation was performed in R. The “dismo” package provides cross-validation using a set of presence and absence locations. The function generates a “ModelEvaluation” object that contains evaluation measures, like AUC, and thresholds which can be used to create the binary maps.

MaxEnt automatically cross-validates the result. Evaluation measures and thresholds were stored in the file “maxentResults.csv” in the output directory that was specified within the modelling process. Binary maps can be created by specifying an appropriate threshold at the “advanced settings” (Figure 18).



☑ Add samples to background
☐ Add all samples to background
☐ Write plot data
☑ Extrapolate
☑ Do clamping
☑ Write output grids
☑ Write plots
☐ Append summary results to maxentResults.csv file
☑ Cache ascii files

Maximum iterations	500
Convergence threshold	0,00001
Adjust sample radius	0
Log file	maxent.log
Default prevalence	0,5
Apply threshold rule	Equal training sensitivity and specificity
Bias file	

Figure 18: Advanced settings in MaxEnt to specify a threshold rule for binary maps.

4 RESULTS

4.1 DISTRIBUTION MAPS

The following section shows the resulting prediction maps. The Figures are structured in the following way:

- Upper left: GLM – continuous prediction map
- Upper right: MaxEnt – continuous distribution map

The maps use the same colour ramp for all species and all modelling methods. The threshold of 0.01 was chosen arbitrarily to distinguish between background and prediction. Values lower than 0.01 are set to transparent.

- Centre left: GLM – binary presences / absence map
- Centre right: MaxEnt – binary presence / absence map

The binary maps only show presence and absence. The threshold for being presence or absence is the value where specificity = sensitivity. Again the same colour was chosen.

- Lower left: atlas map from Niklfeld & Englisch (2004) to compare with known distributions.
- Lower right: Map with the training data showing their distribution within the study area.

4.1.1 CAMPANULA PULLA

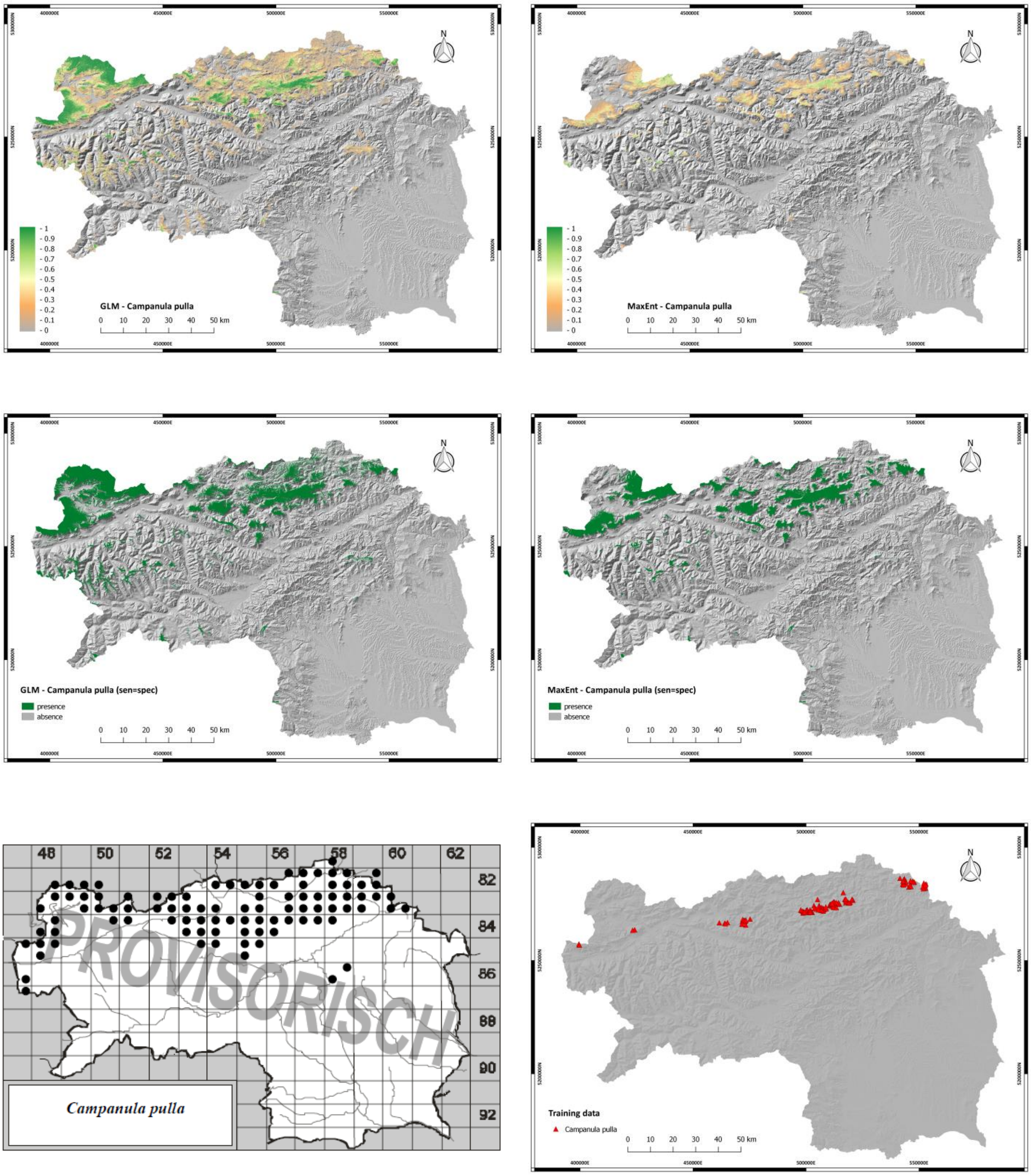


Figure 19: Distribution maps of Campanula pulla. Higher probabilities of occurrence are predicted for the northern limestone Alps. This is also consistent with the known distribution of the species. Furthermore, there are some higher prediction values in the central Alps. The binary maps show larger areas of potential occurrence for GLM compared with MaxEnt. The training data set does not seem to be well distributed throughout the known geographical range. However an east-west distribution is present.

4.1.2 Galium noricum

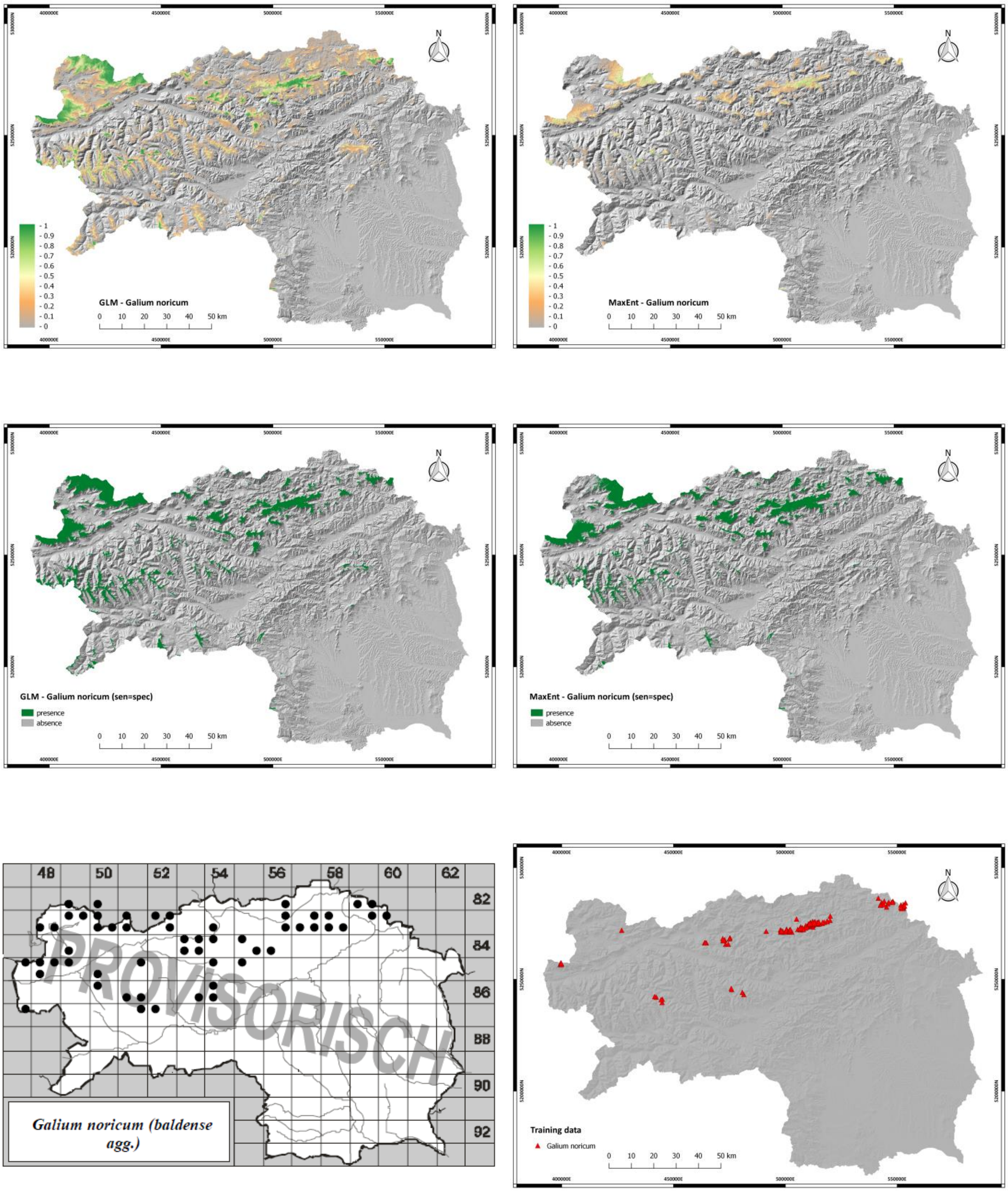


Figure 20: Distribution maps of Galium noricum. The main distribution patterns are similar for both modelling approaches. Again higher probability values were identified mainly in the northern limestone Alps which complies with the atlas map. Several regions within the central Alps could be identified as presence locations. The main east-west distribution of the training record is complemented with some occurrences in the central Alps.

4.1.3 HERACLEUM AUSTRIACUM

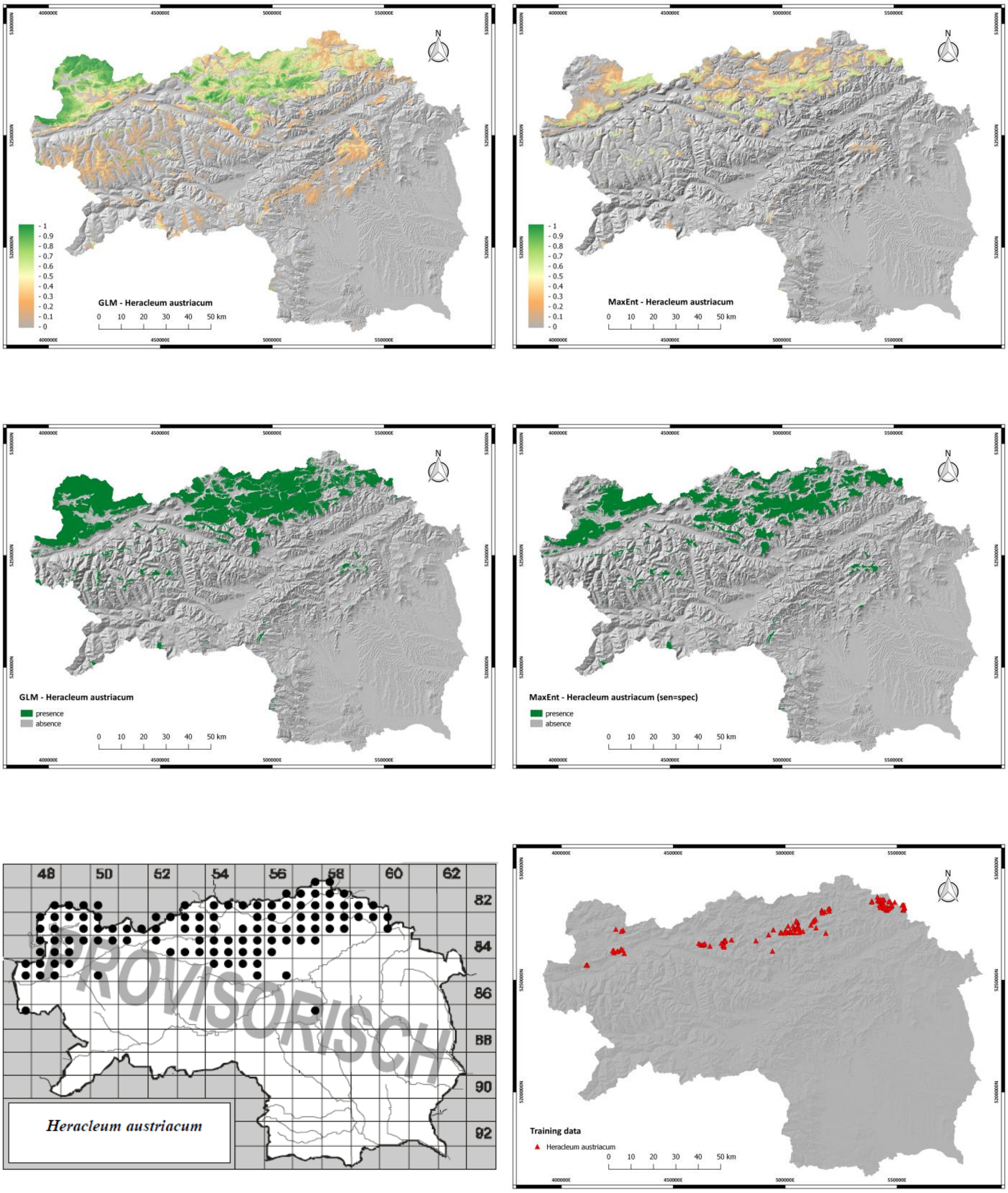


Figure 21: Distribution maps of Heracleum austriacum. The predicted distribution of GLM and MaxEnt identified similar geographic ranges. Higher probability values are concentrated in the northern limestone Alps whereby GLM seems to predict less restrictivly compared with MaxEnt. The binary GLM map shows wide ranges classified as presence location. Within the known geographic range of the species, the training data set is primary restricted to an east-west axis.

4.1.4 PRIMULA CLUSIANA

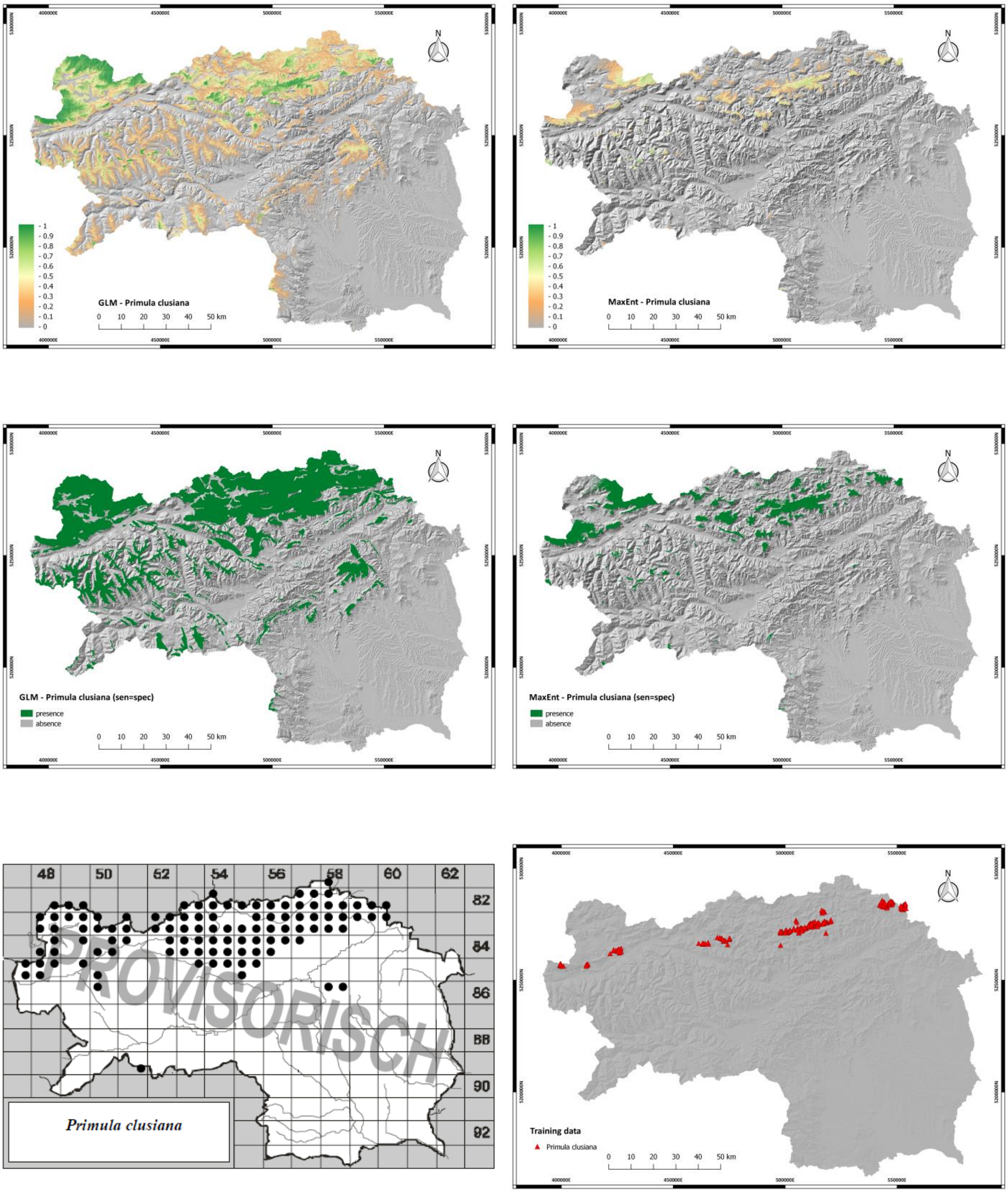


Figure 22: Distribution maps of Primula clusiana. The GLM prediction of Primula clusiana shows high probabilities within the limestone Alps and interestingly a number of regions with medium high values within the central Alps. Accordingly the presence area of the binary map is much larger than the one of the MaxEnt model. The training data set again is concentrated at some areas following an east-west line.

4.1.5 Primula minima

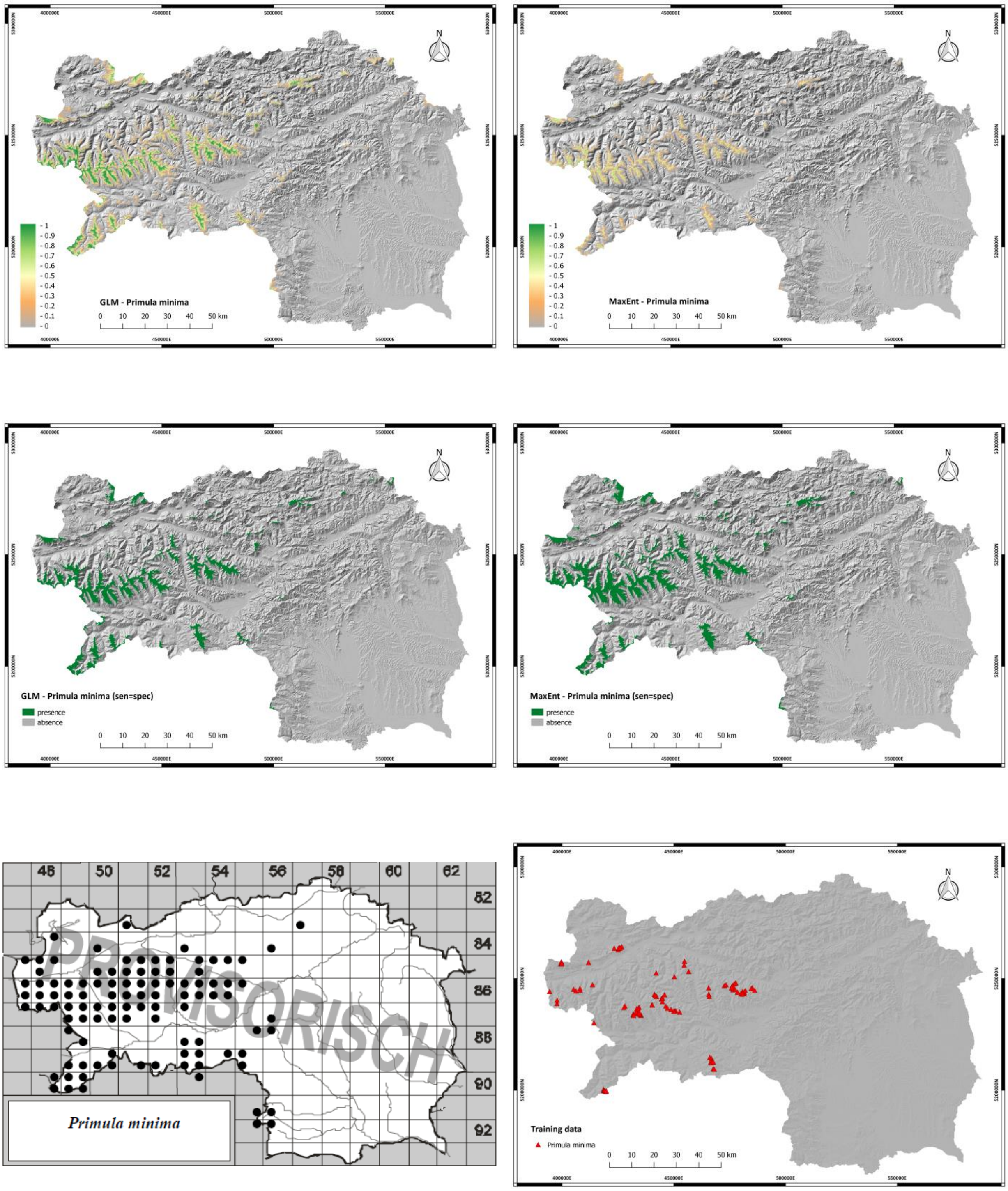


Figure 23: Distribution maps of Primula minima. The prediction maps show a relatively uniform distribution pattern and comply with the known species range. Also the binary maps for both model approaches look similar. The training data set is well distributed across the species geographical range.

4.1.6 VALERIANA CELTICA

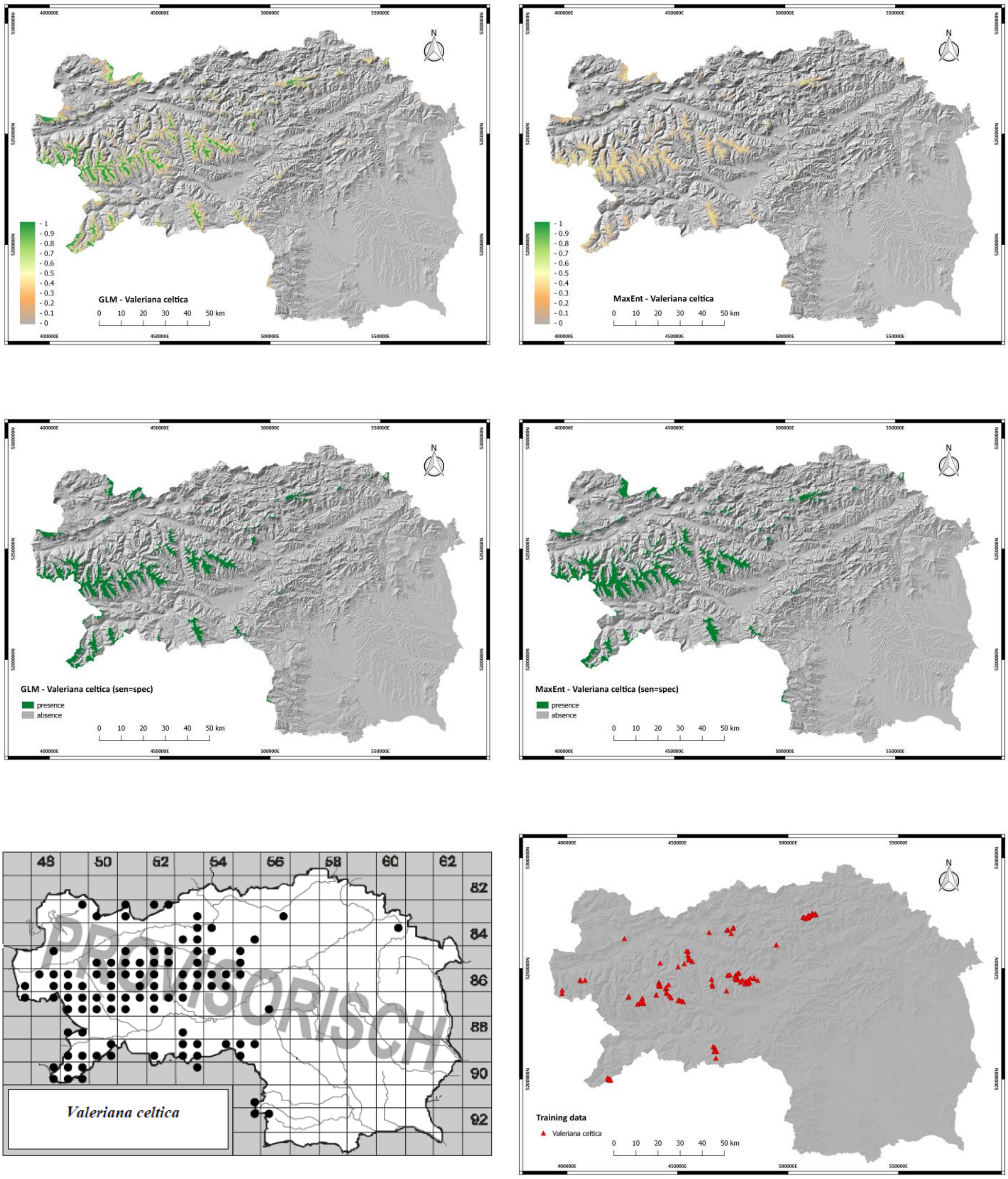


Figure 24: Distribution maps of Valeriana celtica. The predicted distribution looks similar for both modelling approaches. High probability values are concentrated within the known geographic range of Valeriana celtica. Here again, the training data set is well distributed.

4.2 MODEL COMPARISON

The AUC scores for each species distribution model are shown in Table 5. Particularly high values for both modelling approaches and across all species are notable.

AUC values ranges from 0.5 to 1.0 where 0.5 can be interpreted as a random prediction. A value above 0.5 indicates performances that are deemed to be better than random. AUC values of 0.5—0.7 are considered low (poor model performance), values ranging from 0.7—0.9 moderate and above 0.9 high performance (Swets 1988; Franklin 2010).

The AUC scores of both modelling approaches do not vary significantly across the six modelled species. Thus, it is not possible to make a clear ranking between GLM and MaxEnt with the AUC values in Table 5.

Table 5: AUC scores for the species distribution models.

	campull	**galnori**	**primini**	**priclus**	**heraust**	**valcelt**
GLM	0.984	0.992	0.995	0.934	0.951	0.991
MaxEnt	0.986	0.988	0.990	0.985	0.967	0.986

Interestingly, the prediction maps have a different appearance in some parts even though they have similar high AUC scores. Compared with MaxEnt models, GLM seems to be less restrictive with predictions and show more regions with higher values of probability.

Based on a given threshold, the binary maps only distinguish between presence and absence. From these maps it was possible to calculate the total area of presence (area of occupancy). Figure 25 juxtaposes the total area of presence based on the threshold.

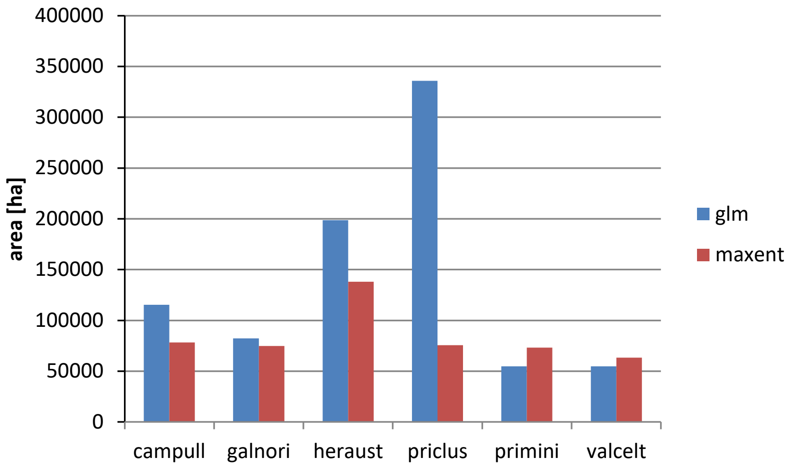


Figure 25: Area of occupancy for all six species derived from the binary maps.

With exception of *Primula minima* and *Valeriana celtica*, the GLM models identified larger areas for species presence than MaxEnt. In the case of *Primula clusiana* the presence area is four times higher compared with MaxEnt.

4.3 SPECIES RESPONSE CURVES

MaxEnt automatically produces species response curves which show how each environmental variable affects the prediction. The following curves show the species response in the model to a specific environmental variable by disregarding all other variables. The influence that each environmental variable has on the occurrence of a species, as well as possible correlations between different variables, can be read from this.

The following figures show the response curves for the three main driving predictors (annual mean temperature, mean summer precipitation and geology) in the MaxEnt models for each of the six species.

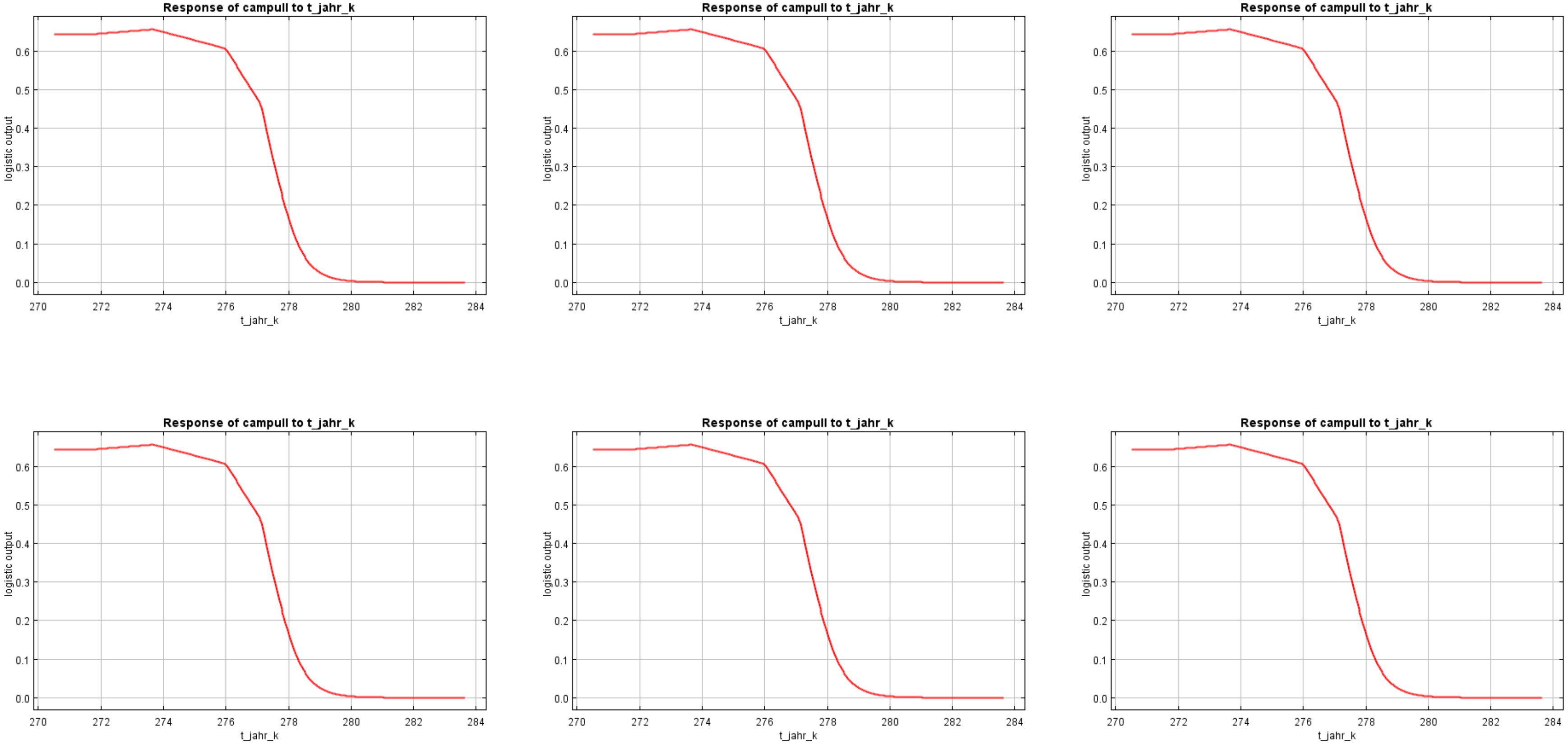


Figure 26: Species response curves to mean annual temperature [Kelvin]. All curves show higher occurrence probabilities at low temperatures.

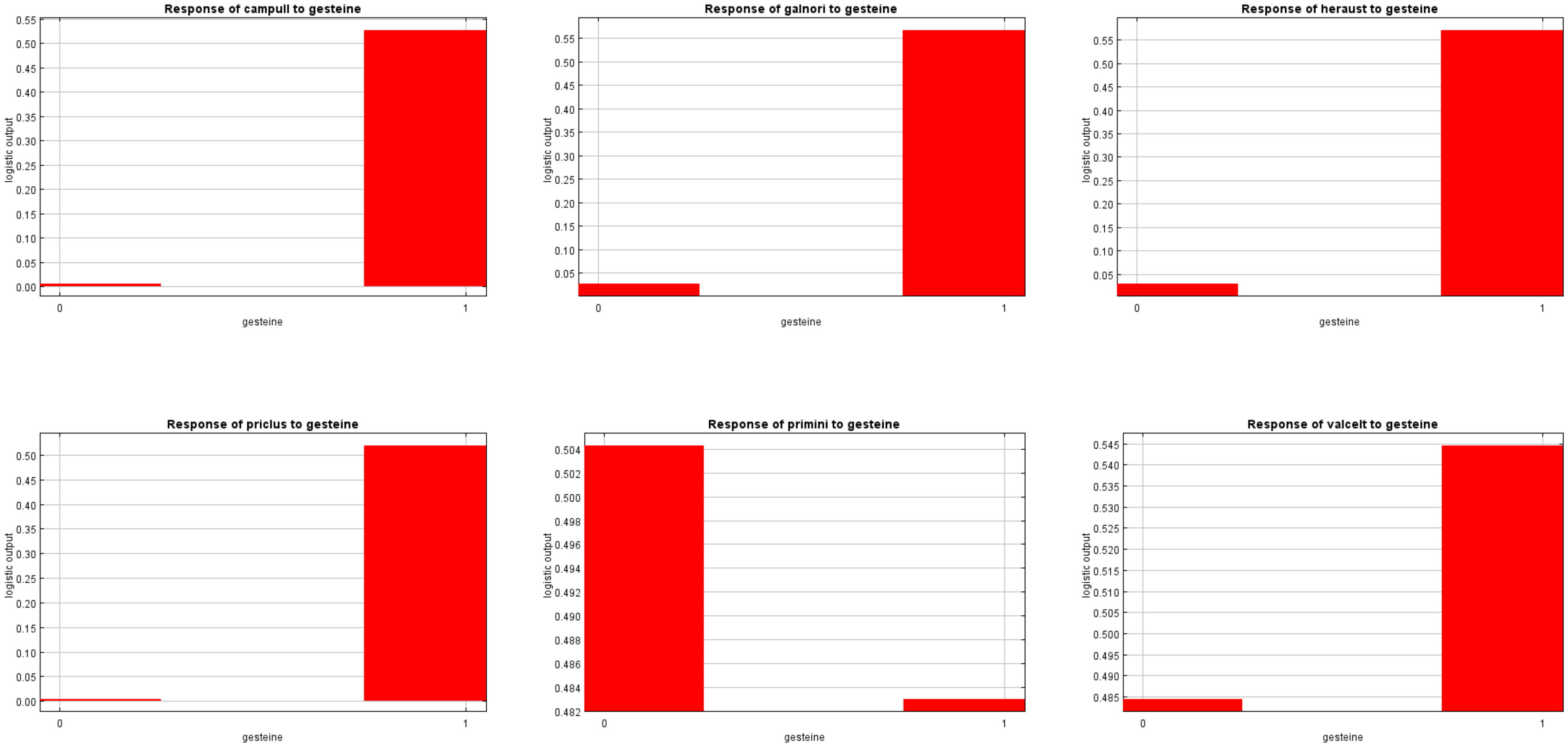


Figure 27: Species response curves to geology. The categorical geological variable is coded with calcareous “1” and non-calcareous”0”. Interestingly, the distribution of Valeriana celtica also appears to be explained by calcareous geology.

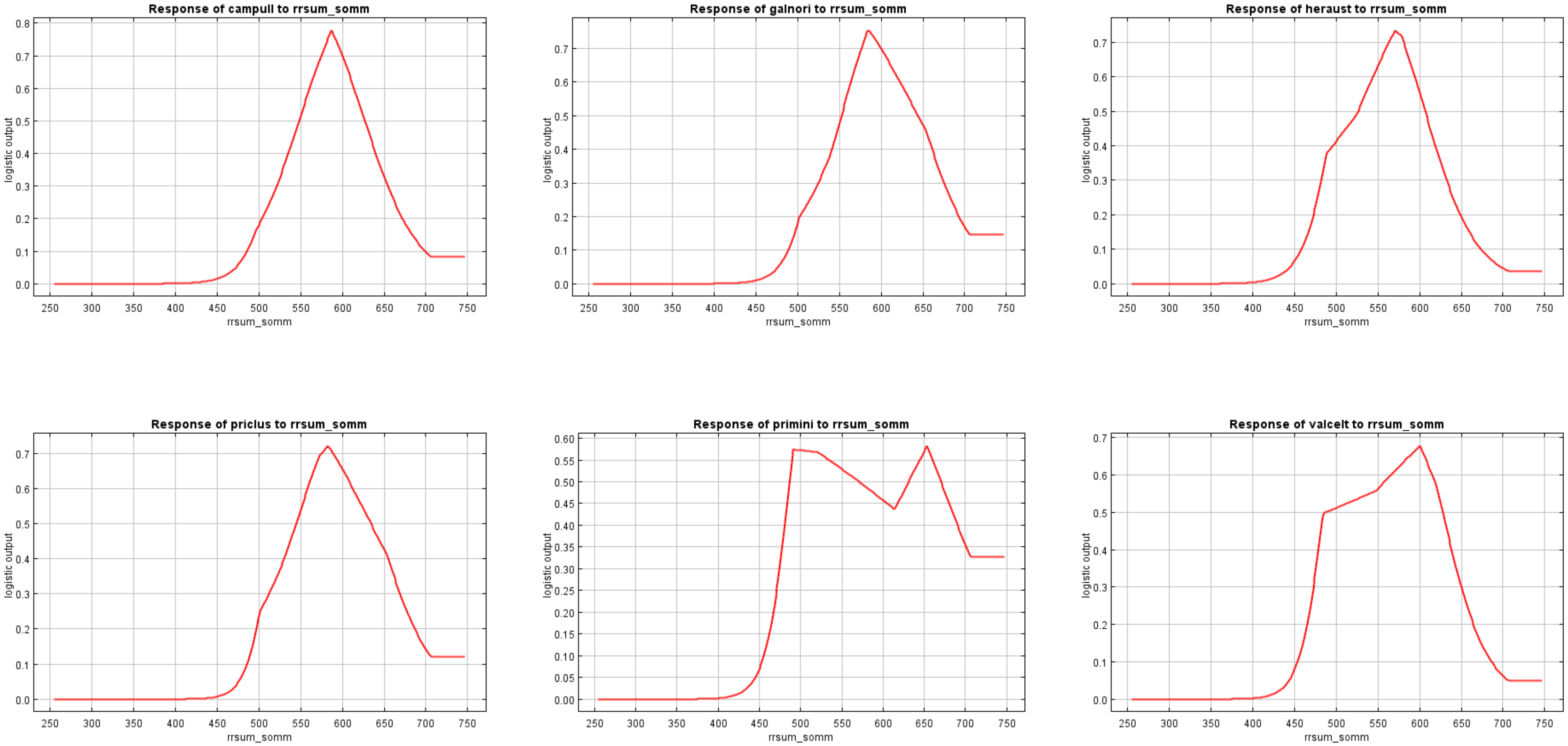


Figure 28: Species response curves to mean summer precipitation. In general, higher occurrence probabilities are correlated with larger values of summer precipitation.

5 DISCUSSION

5.1 POTENTIAL DISTRIBUTION

Both modelling approaches were able to predict the distribution of the six vascular plant species across the Styrian Alps. The models were able to make plausible predictions and comply with known distributions provided by the Styrian distribution atlas. The prediction maps of both approaches show very similar results for one species, whereby GLM models tend to estimate less restrictively. That is, that the method identified more regions to be suitable.

“Completeness” (Kadmon *et al.* 2003) is a measure of how well the species occurrence data covers the environmental space that a species occupies. Especially for extrapolating species in space and time it is important to incorporate well distributed species occurrence data throughout the species’ geographical range and the extent of the environmental gradient within the study area (Franklin 2010). It is only in this way that the observations define the range limits. However, it is possible to predict the habitat suitability of surveyed locations within the study area using only a subset of the species range (Franklin 2010). A set of well distributed (designed) occurrence data would have increased the predictive power of the models. For example, the training samples of *Primula clusiana*, *Heracleum austriacum* and *Campanula pulla* mainly follow an east west axis across the northern limestone Alps. The points are concentrated on a few excursion destinations (e.g. Hochschwab, Rax, Grimming) but are not oriented on geographical ranges of the particular species. Thus, the training data probably do not cover the whole environmental space.

In the case of the calicole alpine species, *Primula clusiana*, *Campanula pulla* and *Heracleum austriacum* it is especially apparent that several new regions within the siliceous central Alps were identified as suitable habitats. These areas are characterized by limestone inclusions that are sparsely distributed over the central

Alps. This pattern can also be seen in the geological layer. Whether this particular species is actually present in the region has yet to be proven in the field.

However, some findings are documented in floristic literature. Occurrences of *Primula clusiana* have been proven to exist at Koralpe, Grebenzen, or Hochlantsch (Maurer *et al.* 1989) and Gumpeneck (Wözer Tauern) (Schneeweiss & Schönswetter 1999). *Campanula pulla* can be found at mount Hochreichart, at Hochlantsch or Gumpeneck (Maurer *et al.* 1989). *Heracleum austriacum* is known from "Steirische Kalkspitze" (Maurer 1996).

The environmental variables used provided good results but the ecological relevance should be interpreted with caution. Table 4 shows the significance of each variable in the GLMs and Figure 17 shows the bar plots of the jackknife test on AUC of MaxEnt. Long blue bars indicate higher explaining contribution. Decreasing green bars indicate loss of high explaining values if this variable is left out. Figure 26 to Figure 28 show the response curves to the main driving predictors (average annual temperature, geology, and average summer precipitation) produced in MaxEnt for each species.

Average annual temperature is the variable that contributes most to the models. It has high significances in the GLMs and seems to hold the most information that is not present in other variables. The response curves show higher response at low temperature values. In fact, average annual temperature is not a particularly relevant variable in ecological sense. The high significance in the models may be explained by the fact that temperatures depend on altitude and the model indicates an optimum range of altitude.

In most cases, the geological factor seems to play a central role (except *Primula minima* and *Valeriana celtica*). It is apparent that this variable has a high explanatory value for all calcicole species. In contrast, *Primula minima* and *Valeriana celtica* are similarily distributed within the central Alps and partially immigrate into the limestone Alps. However, this variable is strongly generalized and only allows for indirect inferences about the substrate conditions. Both species are considered to be pure silicate species and only sparsely occurred on acidified humus deposits. This was the case, for example, with *Valeriana celtica* at Hochschwab plateau (Maurer *et al.* 1989). Interestingly the response of *Valeriana celtica* is slightly shifted to calcareous geology (Figure 27), but comparing the “log output” (y-axis) this shift is quite low.

The optimal water availability for plants during their active time appears to be relevant for their distribution from an ecological point of view. The species response curves show that higher probabilities of occurrence are associated with higher precipitation in the summer months. From a certain value, the curve drops again. However, as already stated in section 3.2, precipitation is a relatively poor proxy for plant water availability. Moreover, precipitation increases with altitude in the Alps and thus may also reflect an altitudinal gradient.

5.2 ARE THE RESULTS COMPARABLE?

To quantify the results a simple method compares the areas of occupancy from the binary maps of both modelling approaches. The threshold was chosen at the values where sensitivity and specificity were equal to one another (equal training sensitivity and specificity). The result is shown in the binary maps in Figure 19 to Figure 24 respectively in Figure 25.

It could be shown that different model approaches (GLM, MaxEnt) have similar distribution patterns, but the same threshold value identifies different areas as potential occurrences. Generally GLM tend to identify larger areas as potential occurrences. However, it is not possible to recommend one approach as “the best method” but it was possible to quantify the differences between the GLM and the MaxEnt results. In the individual case, the choice of the threshold will depend on the question. For ex-situ conservation, the area of occupancy is better defined with a threshold that maximizes the area. Resettlement attempts in restoration ecology, on the other hand, are more likely to be made in areas which are particularly suitable for the respective species. The threshold value is then selected in such a way that the area is reduced to the best regions.

5.3 IS ONE APPROACH PREFERABLE?

GLMs are scientifically proven and have a long tradition in ecological modelling. The application in R requires a little overcoming at the beginning. After a short training the possibilities are very comprehensive.

MaxEnt is very popular these days due to its easy applicability and because it offers a graphical user interface. MaxEnt is also implemented in the “dismo” and “maxnet” package and becomes a more powerful and flexible modelling framework in combination with R.

The MaxEnt algorithm is particularly robust relating to irregular distributed small sample sizes and this makes it especially interesting for modelling rare species.

5.4 PROJECT IMPLEMENTATION

SDMs are of particular importance for a variety of fields in conservation planning. The primer goal is in-situ conservation and thus it is necessary to know where the “hotspots” of a particular species are. Ex-situ conservation can be seen as a supporting measure.

The results have shown that SDMs can assist in localizing special areas of interest. The comparison with available atlas maps seems to coincide with this method. With the help of the prediction maps it is possible to restrict field surveys to particular areas that show higher probability values. And this could contribute significantly to collection success.

However, it is not possible to make a clear statement about population size and thus the expected quantity of a collection. The maps only provide information about how likely it is to find a species within a region based on suitable environmental conditions. The prediction-maps will thus help to find new populations where we can gather seed material for our research and ex-situ collections. In turn, with the new findings of future field surveys it will be possible to improve the models.

6 REFERENCES

Aeschimann, D., Lauber, K. & André Michel, D. (2005). *Flora Alpina: Atlas of 4500 Vascular Plant Species of the Alps : English Introduction*. Haupt.

Austin, M. P. (2002). Spatial prediction of species distribution: An interface between ecological theory and statistical modelling, *Ecological Modelling*, 157(2–3), pp. 101–118.

Cheng, J., Mattiuzzi, M., Sumner, M., Greenberg, J. A., Bevan, A., Shortridge, A. & Hijmans, M. R. J. (2016). Package ' raster '.

Cotto, O., Wessely, J., Georges, D., Klonner, G., Schmid, M., Dullinger, S., Thuiller, W. & Guillaume, F. (2017). A dynamic eco-evolutionary model predicts slow response of alpine plants to climate warming, *Nature Communications*, 8(May), p. 15399.

Dormann, C. F. (2012). *Parametrische Statistik für Ökologen - Verteilungen, maximum likelihood und GLM in R*.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., Mcclean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D. & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance, *Ecography*, 36(1), pp. 027–046.

Dormann, C. F. & Kühn, I. (2009). *Angewandte Statistik für die biologischen Wissenschaften*, *UFZ Umweltforschungszentrum*.

Dormann, C. F., McPherson, M. J., Araújo, B. M., Bivand, R., Bolliger, J., Carl, G., Davies, G. R., Hirzel, A., Jetz, W., Kissling, D. W., Kühn, I., Ohlemüller, R., Peres-Neto, R. P., Reineking, B., Schröder, B., Schurr, M. F. & Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: A review, *Ecography*, 30(5), pp. 609–628.

Dullinger, S., Gattringer, A., Thuiller, W., Moser, D., Zimmermann, N. E., Guisan, A., Willner, W., Plutzar, C., Leitner, M., Mang, T., Caccianiga, M., Dirnböck, T., Ertl, S., Fischer, A., Lenoir, J., Svenning, J.-C., Psomas, A., Schmatz, D. R., Silc, U., Vittoz, P. & Hülber, K. (2012). Extinction debt of high-mountain plants under twenty-first-century climate change, *Nature Climate Change*. Nature Publishing Group, 2(8), pp. 619–622.

Elith, J. & Franklin, J. (2013). Species Distribution Modeling, in Levin, S. A. (ed.) *Encyclopedia of Biodiversity (Second Edition)*. Academic Press, pp. 692–705.

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A.

Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., J. Phillips, S., Richardson, K., Scachetti-Pereira, R., E. Schapire, R., Soberón, J., Williams, S., S. Wisz, M. & E. Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data, *Ecography*, 29(2), pp. 129–151.

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), pp. 43–57.

Essl, F., Staudinger, M., Stöhr, O., Schratt-Ehrendorfer, L., Rabitsch, W. & Niklfeld, H. (2009). Distribution patterns, range size and niche breadth of Austrian endemic plants, *Biological Conservation*. Elsevier Ltd, 142(11), pp. 2547–2558.

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), pp. 38–49.

Fischer, M., Oswald, K. & Wolfgang, A. (2008). *Exkursionsflora für Österreich, Liechtenstein, Südtirol*. 3rd edn. Linz.

Franklin, J. (1995). Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients, *Progress in Physical Geography*, pp. 474–499.

Franklin, J. (2010). *Mapping species distributions. Spatial inference and prediction*. Cambridge Univ. Press (EBC - Ecology, biodiversity and conservation).

Gosch, R. & Berg, C. (2010). Langzeitdiasporenbank steirischer Wildpflanzen am Botanischen Garten Graz, *Mitteilungen des naturwissenschaftlichen Vereins für Steiermark*, 138, pp. 23–28.

Guisan, A. & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models, *Ecology Letters*, 8(9), pp. 993–1009.

Guisan, A. & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology, *Ecological Modelling*, 135(2–3), pp. 147–186.

Harlfinger, O. (2010). Klimaatlas Steiermark : Periode 1971 - 2000, *Studien zum Klimawandel in Österreich*. Wien: Verl. der Österr. Akad. der Wiss.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning*. 2nd edn, *Elements*. 2nd edn. Springer.

Hegel, T. M., Cushman, S. A., Evans, J. & Huettmann, F. (2010). *Spatial complexity, informatics, and wildlife conservation*, *Spatial Complexity, Informatics, and Wildlife Conservation*.

Hijmans, A. R. J., Phillips, S., Leathwick, J., Elith, J. & Hijmans, M. R. J. (2017). Package ‘ dismo ’.

Hijmans, R., Sumner, M., Macqueen, D., Lemon, J. & Brien, J. O. (2016). Package ‘ sp ’.

Hutchinson, G. E. (1957). Concluding remarks., *Cold Spring Harbor Symposia on Quantitative Biology*, 22, pp. 415–427.

John, A., Weisberg, S., Adler, D., Bates, D., Baud-bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Ogle, D., Ripley, B. & Venables, W. (2015). Package ‘car’.

Kadmon, R., Farber, O., Danin, a & Sciences, L. (2003). A systematic analysis of factors affecting the performance of climatic envelope models, *Ecological Applications*, 13(3), pp. 853–867.

Lobo, J. M., Jiménez-valverde, A. & Real, R. (2007). AUC: A misleading measure of the performance of predictive distribution models, *Global Ecology and Biogeography*, 17(2), pp. 145–151.

Lomba, A., Pellissier, L., Randin, C., Vicente, J., Moreira, F., Honrado, J. & Guisan, A. (2010). Overcoming the rare species modelling paradox: A novel hierarchical framework applied to an Iberian endemic plant, *Biological Conservation*. Elsevier Ltd, 143(11), pp. 2647–2657.

Maclean, I. M. D., Hopkins, J. J., Bennie, J., Lawson, C. R. & Wilson, R. J. (2015). Microclimates buffer the responses of plant communities to climate change, *Global Ecology and Biogeography*, 24(11), pp. 1340–1350.

Maurer, W. (1996). Flora der Steiermark : ein Bestimmungsbuch der Farn- und Blütenpflanzen des Landes Steiermark und angrenzender Gebiete am Ostrand der Alpen in zwei Bänden. 1. Farnpflanzen (Pteridophyten) und freikronblättrige Blütenpflanzen (Apetale und Dialypetale). Eching: IHW-Verl.

Maurer, W. (2006). Flora der Steiermark : ein Bestimmungsbuch der Farn- und Blütenpflanzen des Landes Steiermark und angrenzender Gebiete am Ostrand der Alpen in zwei Bänden. 2,1. Verwachsenkronblättrige Blütenpflanzen (Sympetale), *Verwachsenkronblättrige B*. Eching: IHW-Verl.

Maurer, W., Scheuer, C., Baloch, D. & Maurer, W. (1989). Flora der Steiermark : ein Bestimmungsbuch der Farn- und Blütenpflanzen des Landes Steiermark und angrenzender Gebiete am Ostrand der Alpen in zwei Bänden. 2,2. Einkeimblättrige Blütenpflanzen (Monocotyledoneae). Eching: IHW-Verl.

Mod, H. K., Scherrer, D., Luoto, M., Guisan, A. & Scheiner, S. (2016). What we use is not what we know: environmental predictors in plant distribution models, *Journal of Vegetation Science*, 27(6), pp. 1308–1322.

Moser, D., Dullinger, S., Englisch, T., Niklfeld, H., Plutzar, C., Sauberer, N., Zechmeister, H. G. & Grabherr, G. (2005). Environmental determinants of vascular plant species richness in the Austrian Alps, *Journal of Biogeography*, 32(7), pp.

1117–1127.

Müller, J., Berg, C., Détraz-Méroz, J., Fort, N., Lambelet-Haueter, C., Margreiter, V., Mondoni, A., Pagitz, K., Porro, F., Rossi, G., Schwager, P. & Breman, E. (2017). The Alpine Seed Conservation and Research Network - a new initiative to conserve valuable plant species in the European Alps., *Journal of Mountain Scie nce*, 73(2), pp. 251–256.

Niklfeld, H. & Englisch, T. (2004). *Arbeitsatlas zur Farn- und Blütenpflanzenflora der Steiermark*. Graz: Landesmuseum Joanneum.

Oldfield, S. F. (2009). Botanic gardens and the conservation of tree species, *Trends in Plant Science*, 14(11), pp. 581–583.

Ozenda, P. & Borel, J.-L. (2003). The Alpine vegetation of the Alps, in *Alpine Biodiversity in Europe*, pp. 53–64.

Parolo, G., Rossi, G. & Ferrarini, A. (2008). Toward improved species distribution modelling: Arnica montana in the Alps as a case study, *Journal of Applied Ecology*, 45, pp. 1410–1418.

Patsiou, T. S., Conti, E., Zimmermann, N. E., Theodoridis, S. & Randin, C. F. (2014). Topo-climatic microrefugia explain the persistence of a rare endemic plant in the Alps during the last 21 millennia, *Global Change Biology*, 20(7), pp. 2286–2300.

Pebesma, E., Michael, S. & Robert, H. (2017). Package ‘ rgdal ’.

Phillips, S. (2008). A Brief Tutorial on Maxent, *AT&T Research*, pp. 1–38.

Phillips, S. J., Anderson, R. P. & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions, *Ecological Modelling*, 6(2–3), pp. 231–252.

Phillips, S. J. & Dudík, M. (2008). Modeling of species distribution with Maxent: new extensions and a comprehensive evalutation, *Ecograpy*, 31(December 2007), pp. 161–175.

Ripley, W. N. V. and B. D. (2017). Package ‘ MASS ’.

Scherrer, D. & Körner, C. (2011). Topographically controlled thermal-habitat differentiation buffers alpine plant diversity against climate warming, *Journal of Biogeography*, 38(2), pp. 406–416.

Schneeweiss, G. M. & Schönswetter, P. (1999). Feinverbreitung , Ökologie und Gesellschaftsanschluß reliktischer Gefäßpflanzen der Niederen Tauern östlich des Sölkpasses (Steiermark , Österreich), *Stapfia*, 61, pp. 1–242.

Sérgio, C., Figueira, R., Draper, D., Menezes, R. & Sousa, A. J. (2007). Modelling bryophyte distribution based on ecological information for extent of occurrence assessment, *Biological Conservation*, 135(3), pp. 341–351.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems., *Science (New York, N.Y.)*, 240(4857), pp. 1285–1293.

Tichý, L. (2002). JUICE, software for vegetation classification, *Journal of Vegetation Science*, 13, pp. 451–453.

Tobler, W. R. (1970). A Computer Movie Simulation Urban Growth in Detroit Region, *Economic Geography*, 46, pp. 234–240.

Tribsch, A. (2004). Areas of endemism of vascular plants in the Eastern Alps in relation to Pleistocene glaciation, *Journal of Biogeography*, 31(5), pp. 747–760.

Václavík, T., Kupfer, J. A. & Meentemeyer, R. K. (2012). Accounting for multi-scale spatial autocorrelation improves performance of invasive species distribution modelling (iSDM), *Journal of Biogeography*, 39(1), pp. 42–55.

Williams, J. N., Seo, C., 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, *Diversity and Distributions*, 15(4), pp. 565–576.

Willner, W., Berg, C. & Heiselmayer, P. (2012). Austrian Vegetation Database, *Biodiversity & Ecology*, 4, p. 333.

Wyse Jackson, P. & Kennedy, K. (2009). The Global Strategy for Plant Conservation: a challenge and opportunity for the international community, *Trends in Plant Science*, 14(11), pp. 578–580.

Zevenbergen, L. W. & Thorne, C. R. (1987). Quantitative analysis of land surface topography, *Earth Surface Processes and Landforms*, 12(1), pp. 47–56.

7 APPENDIX

7.1 STEPAIC – OUTPUT FOR GLM MODELS

Campabula pulla

```
Start:  AIC=141.23
pb_train ~ easting + gesteine + globre_jr + northing + rrsum_somm + 
    slope + t_jahr_k

             Df Deviance    AIC
- easting     1   125.25 139.25
<none>            125.23 141.23
- rrsum_somm  1   129.25 143.25
- northing    1   130.06 144.06
- globre_jr   1   138.58 152.58
- slope       1   145.40 159.40
- t_jahr_k    1   168.33 182.33
- gesteine    1   188.15 202.15

Step:  AIC=139.25
pb_train ~ gesteine + globre_jr + northing + rrsum_somm + slope + 
    t_jahr_k

             Df Deviance    AIC
<none>            125.25 139.25
+ easting     1   125.23 141.23
- rrsum_somm  1   129.45 141.45
- northing    1   130.08 142.08
- globre_jr   1   138.64 150.64
- slope       1   145.53 157.53
- t_jahr_k    1   168.57 180.57
- gesteine    1   188.27 200.27
```

Galium Noricum

```
Start:  AIC=147.65
pb_train ~ easting + gesteine + globre_jr + northing + rrsum_somm + 
    slope + t_jahr_k

             Df Deviance    AIC
- easting     1   131.79 145.79
- northing    1   132.43 146.43
- rrsum_somm  1   133.08 147.08
<none>            131.65 147.65
- globre_jr   1   133.68 147.68
- slope       1   138.77 152.77
- gesteine    1   169.37 183.37
- t_jahr_k    1   191.62 205.62

Step:  AIC=145.79
pb_train ~ gesteine + globre_jr + northing + rrsum_somm + slope + 
    t_jahr_k

             Df Deviance    AIC
- northing    1   132.58 144.58
- rrsum_somm  1   133.23 145.23
<none>            131.79 145.79
- globre_jr   1   133.81 145.81
+ easting     1   131.65 147.65
- slope       1   138.78 150.78
- gesteine    1   169.37 181.37
- t_jahr_k    1   191.71 203.71

Step:  AIC=144.58
pb_train ~ gesteine + globre_jr + rrsum_somm + slope + t_jahr_k

            Df Deviance    AIC
- rrsum_somm  1   134.36 144.36
- globre_jr   1   134.47 144.47
<none>            132.58 144.58
+ northing    1   131.79 145.79
+ easting     1   132.43 146.43
- slope       1   138.88 148.88
- gesteine    1   171.48 181.48
- t_jahr_k    1   194.44 204.44

Step:  AIC=144.36
pb_train ~ gesteine + globre_jr + slope + t_jahr_k

            Df Deviance    AIC
<none>            134.36 144.36
- globre_jr   1   136.49 144.49
+ rrsum_somm  1   132.58 144.58
+ northing    1   133.23 145.23
+ easting     1   134.20 146.20
- slope       1   140.50 148.50
- gesteine    1   211.47 219.47
- t_jahr_k    1   256.47 264.47
```

Heracleum austriacum

```
Start:  AIC=242.34
pb_train ~ easting + gesteine + globre_jr + northing + rrsum_somm +
    slope + t_jahr_k

            Df Deviance    AIC
- globre_jr   1   226.34 240.34
- northing    1   226.35 240.35
- easting     1   226.39 240.39
<none>            226.34 242.34
- slope       1   228.91 242.91
- t_jahr_k    1   230.23 244.23
- rrsum_somm  1   235.97 249.97
- gesteine    1   283.87 297.87

Step:  AIC=240.34
pb_train ~ easting + gesteine + northing + rrsum_somm + slope +
    t_jahr_k

            Df Deviance    AIC
- northing    1   226.35 238.35
- easting     1   226.39 238.39
<none>            226.34 240.34
+ globre_jr   1   226.34 242.34
- slope       1   230.36 242.36
- t_jahr_k    1   231.70 243.70
- rrsum_somm  1   236.38 248.38
- gesteine    1   283.87 295.87

Step:  AIC=238.35
pb_train ~ easting + gesteine + rrsum_somm + slope + t_jahr_k

             Df Deviance    AIC
- easting     1   226.39 236.39
<none>            226.35 238.35
+ northing    1   226.34 240.34
+ globre_jr   1   226.35 240.35
- slope       1   230.36 240.36
- t_jahr_k    1   231.73 241.73
- rrsum_somm  1   236.89 246.89
- gesteine    1   284.20 294.20

Step:  AIC=236.39
pb_train ~ gesteine + rrsum_somm + slope + t_jahr_k

             Df Deviance    AIC
<none>            226.39 236.39
+ easting     1   226.35 238.35
- slope       1   230.37 238.37
+ northing    1   226.39 238.39
+ globre_jr   1   226.39 238.39
- t_jahr_k    1   231.74 239.74
- rrsum_somm  1   236.89 244.89
- gesteine    1   284.51 292.51
```

Primula clusiana

```
Start:  AIC=218.55
pb_train ~ easting + gesteine + globre_jr + northing + rrsum_somm + 
    slope + t_jahr_k

             Df Deviance    AIC
- slope       1   202.57 216.57
- easting     1   202.58 216.58
- northing    1   203.19 217.19
- globre_jr   1   203.33 217.33
<none>            202.55 218.55
- rrsum_somm  1   211.88 225.88
- t_jahr_k    1   214.66 228.66
- gesteine    1   268.84 282.84

Step:  AIC=216.57
pb_train ~ easting + gesteine + globre_jr + northing + rrsum_somm + 
    t_jahr_k

             Df Deviance    AIC
- easting     1   202.60 214.60
- northing    1   203.68 215.68
- globre_jr   1   204.07 216.07
<none>            202.57 216.57
+ slope       1   202.55 218.55
- rrsum_somm  1   211.89 223.89
- t_jahr_k    1   217.88 229.88
- gesteine    1   270.37 282.37

Step:  AIC=214.6
pb_train ~ gesteine + globre_jr + northing + rrsum_somm + t_jahr_k

             Df Deviance    AIC
```

```
- northing    1   203.69 213.69
- globre_jr   1   204.09 214.09
<none>            202.60 214.60
+ easting     1   202.57 216.57
+ slope       1   202.58 216.58
- rrsum_somm  1   211.90 221.90
- t_jahr_k    1   218.15 228.15
- gesteine    1   270.39 280.39

Step:  AIC=213.69
pb_train ~ gesteine + globre_jr + rrsum_somm + t_jahr_k

           Df Deviance    AIC
- globre_jr   1   204.09 212.09
<none>            203.69 213.69
+ northing    1   202.60 214.60
+ slope       1   203.22 215.22
+ easting     1   203.68 215.68
- rrsum_somm  1   212.36 220.36
- t_jahr_k    1   218.48 226.48
- gesteine    1   276.97 284.97

Step:  AIC=212.09
pb_train ~ gesteine + rrsum_somm + t_jahr_k

           Df Deviance    AIC
<none>            204.09 212.09
+ slope       1   203.37 213.37
+ globre_jr   1   203.69 213.69
+ easting     1   204.07 214.07
+ northing    1   204.09 214.09
- rrsum_somm  1   213.55 219.55
- t_jahr_k    1   218.56 224.56
- gesteine    1   277.10 283.10
```

Primula minima

```
Start:  AIC=127.66
pb_train ~ easting + gesteine + globre_jr + northing + rrsum_somm +
    slope + t_jahr_k

           Df Deviance    AIC
- northing    1   111.79 125.79
- globre_jr   1   111.83 125.83
- easting     1   112.42 126.42
- slope       1   113.60 127.60
<none>            111.66 127.66
- gesteine    1   114.63 128.63
- rrsum_somm  1   117.93 131.93
- t_jahr_k    1   248.06 262.06

Step:  AIC=125.79
pb_train ~ easting + gesteine + globre_jr + rrsum_somm + slope +
    t_jahr_k

           Df Deviance    AIC
- globre_jr   1   111.84 123.84
- easting     1   112.59 124.59
- slope       1   113.62 125.62
<none>            111.79 125.79
- gesteine    1   114.86 126.86
+ northing    1   111.66 127.66
```

```
- rrsum_somm  1   117.93 129.93
- t_jahr_k    1   263.67 275.67

Step:  AIC=123.84
pb_train ~ easting + gesteine + rrsum_somm + slope + t_jahr_k

             Df Deviance    AIC
- easting     1   112.70 122.70
- slope       1   113.68 123.68
<none>            111.84 123.84
- gesteine    1   114.94 124.94
+ globre_jr   1   111.79 125.79
+ northing    1   111.83 125.83
- rrsum_somm  1   117.94 127.94
- t_jahr_k    1   270.46 280.46
```

Valeriana celtica

```
Start:  AIC=118.29
pb_train ~ easting + gesteine + globre_jr + northing + rrsum_somm + 
    slope + t_jahr_k

             Df Deviance    AIC
- easting     1   102.29 116.29
- northing    1   102.89 116.89
- globre_jr   1   103.42 117.42
<none>            102.29 118.29
- slope       1   105.70 119.70
- gesteine    1   107.11 121.11
- rrsum_somm  1   107.70 121.70
- t_jahr_k    1   274.01 288.01

Step:  AIC=116.29
pb_train ~ gesteine + globre_jr + northing + rrsum_somm + slope + 
    t_jahr_k

             Df Deviance    AIC
- northing    1   102.89 114.89
- globre_jr   1   103.42 115.42
<none>            102.29 116.29
- slope       1   105.75 117.75
+ easting     1   102.29 118.29
- gesteine    1   107.12 119.12
- rrsum_somm  1   107.75 119.75
- t_jahr_k    1   275.24 287.24

Step:  AIC=114.89
pb_train ~ gesteine + globre_jr + rrsum_somm + slope + t_jahr_k

             Df Deviance    AIC
- globre_jr   1   103.54 113.54
<none>            102.89 114.89
- slope       1   105.77 115.77
+ northing    1   102.29 116.29
+ easting     1   102.89 116.89
- gesteine    1   107.79 117.79
- rrsum_somm  1   108.30 118.30
- t_jahr_k    1   284.48 294.48

Step:  AIC=113.54
pb_train ~ gesteine + rrsum_somm + slope + t_jahr_k
```

```
             Df Deviance    AIC
<none>            103.54 113.54
- slope       1   105.81 113.81
+ globre_jr   1   102.89 114.89
+ northing    1   103.42 115.42
+ easting     1   103.54 115.54
- gesteine    1   107.96 115.96
- rrsum_somm  1   108.41 116.41
- t_jahr_k    1   306.06 314.06
```