Ecological Modelling 359 (2017) 92–102

Contents lists available at ScienceDirect

Ecological Modelling

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

Research article

Why input matters: Selection of climate data sets for modelling the potential distribution of a treeline species in the Himalayan region

Bobrowski Maria*, Schickhoff Udo

University of Hamburg, Center for Earth System Research and Sustainability, Institute of Geography, Bundesstraße 55, 20146 Hamburg, Germany

ARTICLE INFO

Article history:
Received 17 February 2017
Received in revised form 24 May 2017
Accepted 24 May 2017
Available online 3 June 2017

Keywords:
CHELSA
Betula utilis
Model evaluation
Predictive modeling
WORLDCLIM

ABSTRACT

Betula utilis is a major constituent of alpine treeline ecotones in the western and central Himalayan region. The objective of this study is to analyse for the first time the performance of different climatic predictors in modelling the potential distribution of *B. utilis* in the subalpine and alpine belts of the Himalayan region. Using Generalized Linear Models (GLM) we aim at examining climatic factors controlling the species distribution under current climate conditions. We evaluate the predictive ability of climate data derived from different statistical methods GLMs were created using least correlated bioclimatic variables derived from two different climate data sets: 1) interpolated climate data (i.e. WORLDCLIM; Hijmans et al., 2005), and 2) quasi-mechanistical statistical downscaling (i.e. CHELSA; Karger et al., 2016). Model accuracy was evaluated using threshold-independent (Area Under the Curve) and threshold-dependent (True Skill Statistics) measures. Although there were no significant differences between the models in AUC, we found highly significant differences ($p \leq 0.01$) in TSS. We conclude that models based on variables of CHELSA climate data had higher predictive power, whereas models using WORLDCLIM climate data consistently overpredicted the potential suitable habitat for *B. utilis*.

Although climatic variables of WORLDCLIM are widely used in modelling species distribution, our results suggest to treat them with caution when topographically complex regions like the Himalaya are in focus. Unmindful usage of climatic variables for environmental niche models potentially causes misleading projections.

1. Introduction

The aim of modelling species potential distribution is to characterize suitable habitat conditions, based on climatological, environmental and biotic correlates (Soberón and Nakamura, 2009). The general approach is to link species occurrences with climatic and topographic variables to estimate the species distribution range, since habitat suitability is considerably influenced by the prevailing climate (Pearson and Dawson, 2003). It is assumed that a species occurs within a climatic range determined by its climatic needs within a range of spatial scales (Trivedi et al., 2008).

Within the scope of modelling species niches or distribution, modelling studies face numerous challenges. Not only the choice of modelling algorithm is subject to numerous sources of uncertainties (Elith et al., 2006; Araújo and New, 2007), but also the data used for modelling. Models using presence-absence data have proven to be of great value in predicting species distributions (Guisan et al., 2002; Thuiller et al., 2008), but this data is often not available. Other elements of uncertainties in modelling species distribution are attributed to sample design, sample size, species prevalence, sample resolution, study area extent and the like (for detailed discussion see Franklin, 2009). Further challenges arise from the spatial structure of species occurrence data that may be collinear with environmental data (Araújo and Guisan, 2006; Loiselle et al., 2008; Naimi et al., 2013). Sometimes, areas have been unequally sampled due to differential accessibility of a study area, resulting in occurrences of species with sampling bias. Sampling records often cluster near the centre of climatic conditions under which the species occurs (Loiselle et al., 2008). This leads to species documentations that do not cover the whole range of suitable habitat conditions for respective species. Such geographic sampling bias can lead to sampling bias in environmental space, which represents a major problem for modelling (Veloz, 2009; for the effects of sampling bias on model evaluation see Anderson and Gonzalez, 2011). This holds particularly true for sampling treeline species in remote areas like the Himalayan region. Due to lower accessibility of treeline sites, the number of available sampling plots is sparse, which has a reciprocal effect on prediction performance (Araújo

* Corresponding author.
E-mail address: maria.bobrowski@uni-hamburg.de (B. Maria).

http://dx.doi.org/10.1016/j.ecolmodel.2017.05.021
0304-3800/

et al., 2005). Araújo and Guisan (2006) found that models tend to predict species occupying a narrow niche better than species with a wider niche.

The underlying concept of most modelling studies is the prediction of species distribution ranges using climatic variables. The choice of environmental variables used to model species distributions may result in different distribution maps for the same species (Luoto et al., 2007). Whereas multi-collinearity and spatial auto-correlation of predictors are subject in numerous studies (Dirnböck and Dullinger, 2004; Dormann et al., 2007, 2013; Braunisch et al., 2013), and extensive care is taken in selecting uncorrelated predictor variables, differences in model performance arising from available climate data sets remains largely out of focus in most studies.

Biased climate data can lead to distorted models (Heikkinen et al., 2006). Geographic and environmental biases are contrary to the assumption of many modelling techniques that localities represent a random sample from the area being modelled (Phillips et al., 2006). In many cases, freely available gridded climate data sets do not satisfy the requirements of ecological climate impact studies, and complicate the investigation of climate ecosystem interactions (Soria-Auza et al., 2010).

In the last decade, WORLDCLIM (Hijmans et al., 2005) has been the most prominent global climate data set. Especially in Europe and Northern America, WORLDCLIM shows high accuracy (Hijmans et al., 2005), and is used in numerous biogeographical studies (Elith et al., 2006; Hijmans and Graham, 2006; Broennimann et al., 2012). WORLDCLIM has also been used to model species distributions in the Himalayan region (Forrest et al., 2012; Liu et al., 2017), the accuracy, however, needs to be evaluated. Bobrowski et al. (2017) pointed out some drawbacks, related to the usage of WORLDCLIM.

WORLDCLIM represents a simple interpolated climate data set, which regionalizes monthly observations of precipitation and temperature based on a weighted linear regression approach, using latitude, longitude and elevation as predictor variables. Despite the high spatial raster resolution (i.e. 1×1 km), WORLDCLIM ignores atmospheric processes at local scale which are essential for the formation of site-specific topoclimatic conditions in high mountain environments. Many studies show that local-scale atmospheric conditions are highly influenced by the underlying terrain. Anisotropic heating at different slope positions as well as cold air drainage and pooling in mountain valleys during autochthonous weather conditions result in a complex temperature pattern, which distinctly modifies the distribution of plant communities (Bobrowski et al., 2017). The spatial pattern of precipitation is affected by wind- and leeward slope positions, with hyper-humid climate conditions at the southern declivity of the Himalayan range and semi-arid to arid conditions in the *trans*-Himalayan valleys.

Since 2016, a new fine-scale (i.e. 1×1 km), long-term climate raster data set with global coverage called CHELSA (Climatologies at high resolution for the earth' land surface areas) is available (Karger et al., 2016). CHELSA was compared and evaluated with three climate data sets (i.a. WORLDCLIM), and showed similar performance for temperature, but higher performance for prediction of orographic precipitation patterns (Karger et al., 2016). Both climate data sets use the same raw data to produce the same bioclimatic raster-layers. However, CHELSA represents the first global climate data set based on statistical downscaling, whereas WORLDCLIM is based on interpolation.

To date, there are only very few studies aiming at comparing and evaluating modelling results obtained by different (e.g., climate data sets (comparison of SAGA and WORLDCLIM in Soria-Auza et al., 2010 using Böhner, 2006 and Hijmans et al., 2005). Comparative studies, which evaluate the performance of ecological niche models using different global climate data sets for modelling the potential distribution of Himalayan treeline tree species' or other Himalayan vascular plant species' do not exist. We selected the treeline-forming species *Betula utilis* as a target species since an improved accuracy in modelling the current distribution is a precondition for a more precise modelling of potential range expansions of treeline trees under climate change conditions (Schickhoff et al., 2015).

In order to investigate the impact of each climate data set we compared the predicted current distribution of *Betula utilis* in the Himalayan region. We applied Generalized Linear Models, using each climate data set respectively, to model the distribution range and compare and evaluate projected distribution range maps. We hypothesized that there will be discrepancies in the predictions of the two climate data sets. We assume a higher prediction accuracy of CHELSA because of its capability to reflect mountain-specific climatic conditions, in particular in terms of precipitation-related variables.

2. Material and methods

2.1. Study area and species data

The Himalayan mountain system is located between the Tibetan Highland in the north and the Indo-Gangetic plains in the south, and extends from Afghanistan in the northwest (c. 36°N and 70°E) to Yunnan in the southeast (c. 26°N and 100°E). It is a vast mountain region, covering an area of more than 1.000.000 km^2, with a length of c. 3000 km (Pakistan to SW China) and a maximum width of 400 km.

The Himalayan mountains show a distinct three-dimensional geoecological differentiation, with complex topography and high variation of climatic and edaphic conditions (Miehe, 2015). The climate ranges from tropical in the Indo-Gangetic plains to permanent ice and snow at highest elevations, and from more continental in the NW to more oceanic in the SE (Troll, 1972; Zurick and Pacheco, 2006). The distribution range of *Betula utilis* extends across the Himalayan range from Afghanistan to SW China, with the total elevational range extending from 2700 to 4500 m (Polunin and Stainton, 1984). *B. utilis* was selected as a study species due to its status as a principal broadleaved treeline species in the western and central Himalaya. The Himalayan birch mainly grows on north-facing slopes in shady locations. In the NW Himalaya, *B. utilis* is widely distributed between 3100 and 3700 m, while the range shifts to higher elevations towards the E Himalaya (mainly between 3800 and 4300 m; own database). *B. utilis* grows in mixed forests with conifers and rhododendrons and forms a narrow forest belt between coniferous forests below and a krummholz belt above (for associated tree species see Schickhoff, 2005; Miehe et al., 2015). Pure birch stands with *Rhododendron campanulatum* and *Sorbus microphylla* in the understory are often found at the uppermost limit of subalpine forests (Schickhoff et al., 2015).

Presence-only occurrence data of *B. utilis* were accessed via the Global Biodiversity Information Facility (GBIF, 2016). The database hosts 215 geo-referenced records (1970–2016) without any known coordinate issues for our study region. 202 records were added from a database compiled from a literature survey (Schickhoff, 2005; unpublished data). Additionally, we extracted 327 records from freely available satellite images (GoogleEarth™, ver. 7.1.1.1888, Google, 2015) and added them to the dataset. Extractions from GoogleEarth have been shown to be useful in global treeline research (Paulsen and Körner, 2014; Irl et al., 2015). These occurrence localities were validated through expert knowledge, obtained from numerous field visits in the Himalayan region. To reduce spatial auto-correlation, only one occurrence point per grid cell (i.e. 1×1 km) was considered, resulting in 590 occurrence points of *B.*