Journal of
Plant Ecology

VOLUME 5, NUMBER 3,
PAGES 337–345

JULY 2012

doi:10.1093/jpe/rtr049

Advanced Access published
on 11 January 2012

available online at
www.jpe.oxfordjournals.org

Impacts of predictor variables and species models on simulating *Tamarix ramosissima* distribution in Tarim Basin, northwestern China

Qiang Zhang[1,2] *and Xinshi Zhang*[1,*]

[1] *State Key Laboratory of Vegetation and Environmental Change, Institute of Botany, Chinese Academy of Sciences, 20 Nanxincun, Xiangshan, Beijing 100093, China*
[2] *Graduate University of Chinese Academy of Sciences, 19A Yuquanlu, Beijing 100049, China*
*Correspondence address. State Key Laboratory of Vegetation and Environmental Change, Institute of Botany, Chinese Academy of Sciences, 20 Nanxincun, Xiangshan, Beijing 100093, China. Tel: +86-10-58808555; Fax: +86-10-58213037; E-mail: xinshiz@yahoo.com

Abstract

Aims

Preserving and restoring *Tamarix ramosissima* is urgently required in the Tarim Basin, Northwest China. Using species distribution models to predict the biogeographical distribution of species is regularly used in conservation and other management activities. However, the uncertainty in the data and models inevitably reduces their prediction power. The major purpose of this study is to assess the impacts of predictor variables and species distribution models on simulating *T. ramosissima* distribution, to explore the relationships between predictor variables and species distribution models and to model the potential distribution of *T. ramosissima* in this basin.

Methods

Three models—the generalized linear model (GLM), classification and regression tree (CART) and Random Forests—were selected and were processed on the BIOMOD platform. The presence/absence data of *T. ramosissima* in the Tarim Basin, which were calculated from vegetation maps, were used as response variables. Climate, soil and digital elevation model (DEM) data variables were divided into four datasets and then used as predictors. The four datasets were (i) climate variables, (ii) soil, climate and DEM variables, (iii) principal component analysis (PCA)-based climate variables and (iv) PCA-based soil, climate and DEM variables.

Important Findings

The results indicate that predictive variables for species distribution models should be chosen carefully, because too many predictors can reduce the prediction power. The effectiveness of using PCA to reduce the correlation among predictors and enhance the modelling power depends on the chosen predictor variables and models. Our results implied that it is better to reduce the correlating predictors before model processing. The Random Forests model was more precise than the GLM and CART models. The best model for *T. ramosissima* was the Random Forests model with climate predictors alone. Soil variables considered in this study could not significantly improve the model's prediction accuracy for *T. ramosissima*. The potential distribution area of *T. ramosissima* in the Tarim Basin is $\sim 3.57 \times 10^4$ km^2, which has the potential to mitigate global warming and produce bioenergy through restoring *T. ramosissima* in the Tarim Basin.

Keywords: species distribution model • *Tamarix ramosissima* • generalized linear models • classification and regression trees • RandomForest

Received: 9 June 2011 Revised: 30 November 2011 Accepted: 3 December 2011

INTRODUCTION

The modelling of vegetation and species distributions based on their relationship with environmental variables is important to many management activities. Examples include identifying suitable areas for reintroduction of species and restoration of vegetation (Austin 2007; Elith and Leathwick 2009; Guisan *et al.* 1998; Guisan and Zimmermann 2000), evaluating the risk of non-native species in new environments (Elith and Leathwick 2009; Evangelista *et al.* 2009; Feagin 2005; Guisan and Harrell 2000; Ibanez *et al.* 2009; Jones *et al.* 2010) and estimating the magnitude of biological responses to climate change

(Abbott and Le Maitre 2010; Barry and Elith 2006; Ferrier 2002; Hamann and Wang 2006; Mckenney *et al.* 2007; Pearson and Dawson 2003; Randin *et al.* 2009; Retuerto and Carballeira 2004; Yates *et al.* 2010). Dozens of models are available for describing the relationship between environmental variables and species distributions (Guisan and Harrell 2000), among which generalized linear models (GLM) (McCullagh and Nelder 1989), classification and regression trees (CART) (Breiman *et al.* 1984; De'ath and Fabricius 2000) and Random Forests (Araujo and New 2007; Breiman 2001; Garzon *et al.* 2006; Lawler *et al.* 2006; Peters *et al.* 2009; Prasad *et al.* 2006) are frequently applied. Because of the complexity of the nature, distribution modelling results inevitably contain some degree of uncertainty (Barry and Elith 2006). One of the major sources of the uncertainty in species distribution models comes from the limitations of input data, specifically the spatial and temporal under-representation of observations, measurement and systematic errors in observations, missing key environmental variables and the collinearity and spatial autocorrelation of environmental variables (Barry and Elith 2006; Ray and Burgman 2006). Furthermore, distribution modelling techniques introduce uncertainty by their inability to capture the entire complexity of ecological processes associated with vegetation distributions. There are few comparative works where more than two statistical methods are applied to the same dataset (Elith and Leathwick 2009; Guisan and Harrell 2000), although the assessment of uncertainty receives more and more attention in ecological modelling studies (Larssen *et al.* 2007; Peters *et al.* 2009; Phillips and Marks 1996; van Horssen *et al.* 2002; Van Niel and Austin 2007).

Tamarix ramosissima, commonly known as Saltcedar, plays important roles in desert ecosystems and serves as an important source of fuel and forage for local populations (Gries *et al.* 2005; Yang *et al.* 2004). It is native to Asia and distributes widely, while it is a notorious invasive species in North America (Cleverly *et al.* 1997; Stromberg *et al.* 2007), Argentina, Australia and South Africa. *Tamarix ramosissima* was once widely distributed throughout the Tarim Basin in Northwest China. The abundance of *Tamarix* vegetation has declined dramatically in the last decades because of the excessive deforestation and exploitation of groundwater, which has exacerbated the effects of desertification in the Tarim Basin (Liu 1995). As a halophyte with high salt tolerance and high resistance to drought, wind erosion and sand burial, *T. ramosissima* has been widely used in desertification control in China (Liu 1995). In addition, *T. ramosissima* is considered as a promising bioenergy plant that can be grown in desert regions (Abideen *et al.* 2011; Eshel *et al.* 2010; Feng 2008; Li *et al.* 2009; Tang *et al.* 2010). In recent years, as the major host plant of the parasitic *Cistanche* spp., which is a profitable medicinal plant, extensive efforts to restore *T. ramosissima* were commenced in Xinjiang, China. Thus, the restoration of *Tamarix* vegetation is not only an urgent requirement for desertification control but also an important demand for economic interests. An important prerequisite for successfully restoring vegetation is the knowledge of suitable distribution habitats for species of interest. Generally, the application of species distribution models is an effective approach to identify the potential distribution habitat for successful restoration.

The purpose of this paper was (i) to assess the effects of chosen predictor variables and their principal component analysis (PCA) axis on the performance of GLM, CART and Random Forests models, (ii) to identify the appropriate environmental variables for modelling *T. ramosissima* distribution, (iii) to compare the performance of GLM, CART and Random Forests on modelling *T. ramosissima* distribution and (iv) to model the potential distribution of *T. ramosissima* in the Tarim Basin.

MATERIALS AND METHODS

Site description

The Tarim Basin is the largest basin in China and one of the largest internal drainage basins in the world, and it is surrounded by alpine mountain, where the Pamir lies to the west, the Tianshan Mountains to the north and the Kunlun Mountains to the south (Fig. 1). Diluvial–alluvial plains make up the foreland of the basin. The biggest shifting desert in the world—the Taklamakan desert—is located at the centre of the basin, and the Lop Nur plain, which is the lowest region of the basin, makes up the easet basin. The zonal soil in the basin is blown desert soil, and the azonal soil contains meadow, salt and aeolian soil. The climate is extremely continental, with cold, dry winters and hot, dry summers. The mean annual temperature (MAT) is ~11.1°C. The lowest annual air temperature is generally around −20°C to −30°C, while the highest annual air temperature can reach 47.6°C. The annual potential evaporation is ~2 600 mm, and the annual precipitation is 50–70 mm in the north and 15–30 mm in the south. As northwest winds prevail in the western Taklamakan and northeast winds in the east, aeolic sediments constantly accumulate at the southern margin (Zhang 2011; Zhang and Runge 2006).

Data collection

The presence/absence data of *T. ramosissima* vegetation were calculated from the newest 1:1 000 000 vegetation map of China edited by Zhang (2008) and used as the response variable in models. Digital elevation model (DEM) data and 12 climatic and 14 edaphic environmental parameters were used as predictors to establish the model.

The DEM data and soil data were obtained from Void-filled seamless SRTM data V1 (International Centre for Tropical Agriculture (CIAT), 2004, CGIAR-CSI SRTM 90m Database, http://srtm.csi.cgiar.org) and the Harmonized World Soil Database produced by IIASA (2008), respectively. Both datasets were made available by WESTDC (Environmental and Ecological Science Data Center for West China, National Natural Science Foundation of China, http://westdc.westgis.ac.cn) and were resampled to 10 × 10 km from 1 × 1 km resolution. The original soil data were produced from a series of soil maps

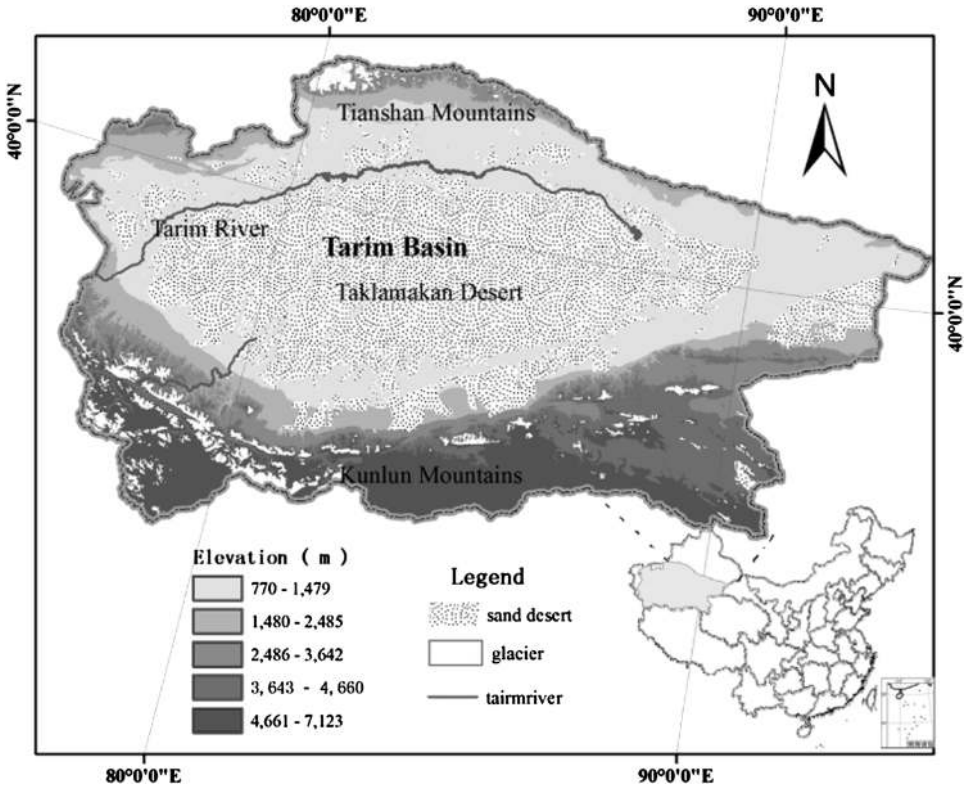

Figure 1: sketch map of the study area.

covering the extent of China at a scale of 1:1 million based on the Second National Soil Survey of China and were transformed to a digital format by the Institute of Soil Science, Chinese Academy of Sciences, Nanjing. The edaphic factors used as predictors in this study included the content of gravel, sand, silt and clay, organic carbon, pH, Electrical Conductivity (ECE) and bulk density within topsoil (0–30 cm) and subsoil (30–100 cm). Among these environment variables, soil organic matter is the main nutrient source for plants, and soil pH and ECE exert direct physiological limitations on plants, while elevation, soil bulk density, soil gravel, sand, slit and clay content would affect the availability of nutrients, water and heat to plant.

Climatic factors used in the models were MAT, mean annual precipitation (MAP), CI (Kira cold index), WI (Kira warm index), mean temperature in the growing season from April to September (GST), mean precipitation in the growing season from April to September (GSP), PET (potential evapotranspiration from the United Nation's Food and Agriculture Organization, Allen *et al.* 1998), range of annual temperature (ATR), Holdridge's biotemperature (BT), mean temperatures in July (JulT), mean temperature in January (JanT) and the Arid index (AI, AI = PET/AMP). Among the selected climate variables, MAT, GST, ATR, BT, WI and CI reflect the heat condition and energy supply for plant growth and development; JulT and JanT reflect the extreme temperatures that plants can endure and survive; MAP, GST, PET and AI reflect the water supply and the degree of dryness tolerable for plant growth and survival. Because there are only 14 climate stations in the Tarim Basin, these indexes were calculated by interpolating data recorded at 752 standard climate stations over China with 10 × 10 km resolution employing the kriging method. The resampling and interpolation of the spatial data were processed with Arcgis9.3 software.

Statistical models

The GLM was introduced by Austin *et al.* (1984) to model the presence/absence data of the tree species. The GLM method provides a less restrictive form than classic multiple regressions by providing error distributions for the dependent variable other than normal and non-constant variance functions (McCullagh and Nelder 1989). If the response with a predictor variable is not linear, then a transformation can be included; polynomial terms are allowed for the simulation of skewed and bimodal responses (Guisan *et al.* 1999), β functions (Austin and Gaywood 1994) or hierarchical sets of models (Huisman *et al.* 1993). The associated shortcoming of GLM is that the nature of the relationship between species and environmental gradients has to be known *a priori*. Furthermore, the GLM cannot deal with complex response curves (Yee and Mitchell 1991).

CART was developed by Breiman *et al.* (1984). Rather than trying to identify and model a general relationship between predictor variables and responses, CART recursively partitions the multidimensional space defined by the predictor variables into zones that are as homogeneous as possible in terms of

response. The tree is built by repeatedly splitting the data, defined by a simple rule based on a single explanatory variable. At each split, the data are partitioned into two exclusive groups, each of which is as homogeneous as possible (Thuiller 2003). CART is less commonly used than GLM methods but is accurate and useful to describe hierarchical interactions between species (Franklin 1998; Thuiller *et al.* 2003). The main drawback of CART model is that the generated models can be extremely complex and difficult to interpret when used to predict organism distributions, with more than just a handful of predictor variables or cases to classify (Muñoz and Felicísimo 2004).

Random Forests modelling is an ensemble learning technique that generates many classification trees that are aggregated, based on majority voting, to classify (Breiman 2001; Breiman *et al.* 1984). Bootstrap samples are drawn to construct multiple trees, each tree is grown with a randomized subset of total number of predictors and a large number of trees are grown. Observations in the original dataset that do not occur in a bootstrap sample are called out-of-bag observations (OOB) and can be used to calculate an unbiased error rate and variable importance, eliminating the need for a test set or cross-validation. The trees are grown to maximum size without pruning, and each is used to predict the OOB. The predicted class of an observation is calculated by majority vote of the OOB for that observation. Random Forests produces a limiting value of the generalization error, which means that no overfitting is possible, a very useful feature for prediction (Breiman 2001; Prasad *et al.* 2006).

CART, GLM and Random Forests models were established to predict the presence/absence of *T.ramosissma* using four datasets. The four datasets were (i) climatic variables, (ii) climatic variables, edaphic variables and the DEM, (iii) PCA axes of climatic variables, (iv) PCA axes of climatic variables, edaphic variables and the DEM. The presence/absence data of *T.ramosissma* were randomly divided into two groups, which were then used to split four environmental datasets into two groups, where 80% of data were used to build the model and 20% of data were used to calibrate and evaluate the model. Each model was constructed using the BIOMOD platform (Thuiller 2003; Thuiller *et al.* 2009). The stepwise procedure of the GLM based on Akaike's information criterion. The number of repetitions in BIOMOD was set to three. The area under the receiver operating characteristic (ROC) curve (AUC) was used to evaluate the model performance. The other options of BIOMOD were set to default. The differences among the three models was tested by one-way ANOVA using R 2.12.1 software (R Development Core Team 2010).

RESULTS

Effects of chosen environmental variables and PCA management of predictor variables on the model performance

The effects of using different environment variables on the model are illustrated in Fig. 2. In the case of the CART model, the performance was better for the dataset of only climate variables than for the dataset of climate, soil and DEM data. However, this was not the case when using PCA-based data. In the case of the GLM, the performance with climate, soil and DEM data was better than that with only climate variables, but the use of PCA-based data did not result in significantly better performance than the use of the original data. In the case of the Random Forests model, there was no significant difference in performances between only with climate variables and with climate, soil and DEM data, while the performance achieved by using the PCA-based dataset of climate, soil and DEM data was better than that using the PCA-based dataset of only climate variables. Additionally, the original datasets performed better than that of PCA-based data did. However, none of the above described different was significant.

Best model for predicting the potential distribution of *T. ramosissima*

The performances of the different models are shown in Fig. 3 and the average AUC for each *T. ramosissima* model can be seen in Table 1. In light of AUC, the performance of Random Forests model was the best and followed by the GLM and then the CART model. When comparing different models within different datasets, the Random Forests moldel outperformed than GLM and CART with CART having the lowest performance (Table 1). The Random Forests model built by the dataset of climate variables had the highest AUC (0.956). Thus, this model was considered as the best model for predicting the potential distribution of *T. ramosissima*.

Potential distribution of *T. ramosissma* in the Tarim Basin

The potential distribution of *T. ramosissima* in the Tarim Basin was predicted by the Random Forests model only with climate data (Fig. 4). The predicted result was close to the original distribution. Compared with the actual distribution of *T. ramosissma* from the vegetation map to the predicted distribution area from the model, for 19.6% of the predicted inhabited area, *T. ramosissima* was not reported in datasets. For 15% of the inhabited area reported in datasets, the model failed to predict the presence of the plant. The predicted potential distribution area of *T. ramosissima* was $\sim 3.57 \times 10^4$ km^2. *Tamarix ramosissima* is distributed mainly around the borders of the Taklamakan desert and along the Tarim River, especially the northern Tarim Basin.

DISCUSSION

The missing of key predictor variables is considered to be the main source of uncertainty (Barry and Elith 2006; Guisan and Harrell 2000). For all the models, results in the current study showed that the performance of the dataset with climate, soil and the DEM variables was better than that of the dataset with climate variables alone (Fig. 1). Previous researchers have shown that it is not wise to use too many predictor variables

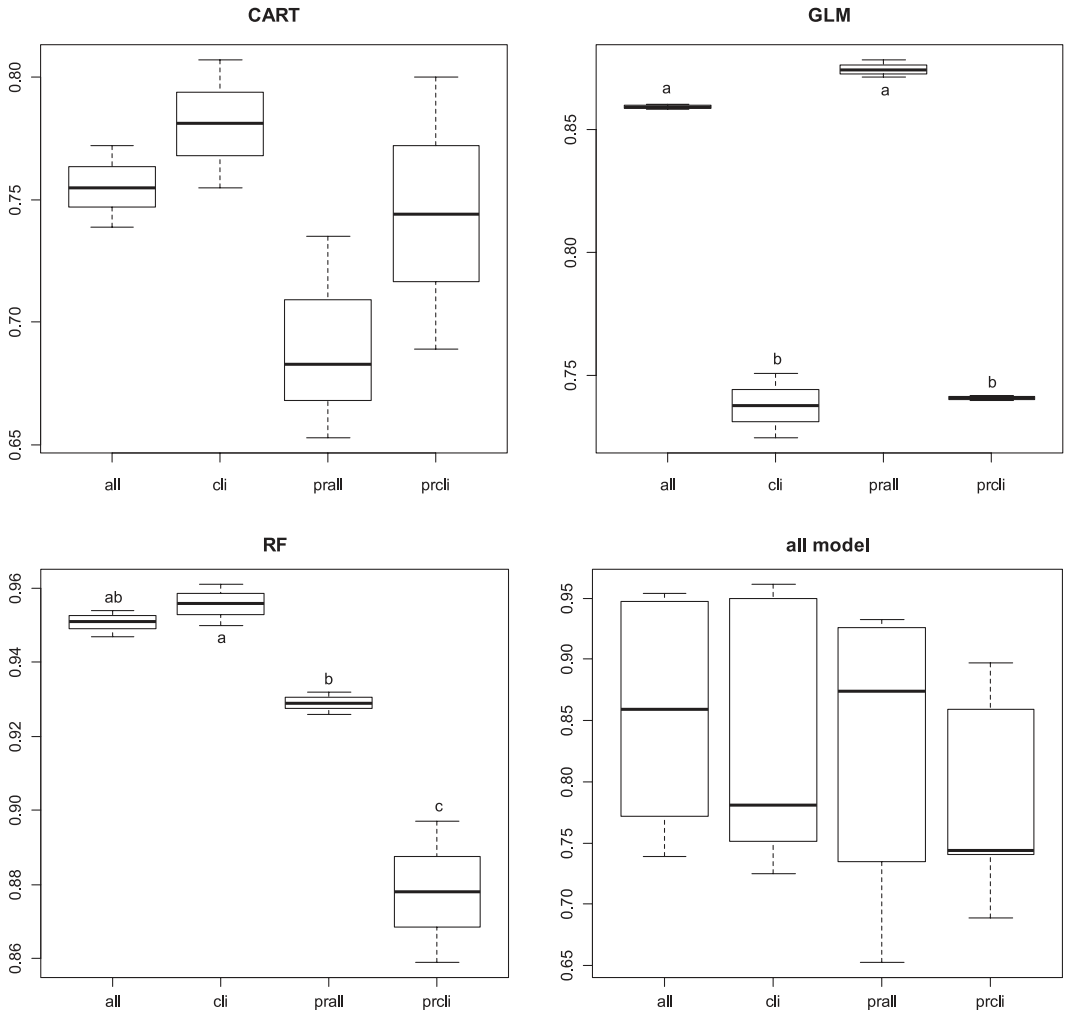

Figure 2: effects of the choice of the predictor variables and the PCA management of predictor variables on the model. The abbreviations in this figure are the same as in Table 1.

in model (Barry and Elith 2006; Guisan and Harrell 2000). Our results indicate that sometimes a greater number of predictor variables result in poor performance of the model (Fig. 1). More attention should be paid to the selection of predictors before establishing the model.

Tamarix ramosissima as an azonal vegetation is affected by groundwater, flood inundation and soil salt (Gries *et al.* 2003; Yang *et al.* 2004). However, data on the groundwater table are scarce. In this study, the study area spanned the Tarim Basin plain, which is flat, and the sedimentary characteristic generally is consistent because the whole basin sits on the Tarim platform. Therefore, the altitude could be considered to represent the groundwater table here. Since *T. ramosissima* could distribute in sand, loam habitat and even in the infertile Gobi desert and has high salt resistance, therefore, none of the soil particles' composition, salinity and nutrient could be the main factors to limit its distribution. In addition, the resolution of soil data was low. Consequently, the edaphic factors in this study did not improve the prediction accuracy.

PCA is generally used to avoid the collinearity of correlated predictor variables (Dormann *et al.* 2008; Elith *et al.* 2011; Mellin *et al.* 2010; Rotenberry *et al.* 2006; Townsend *et al.* 2007) and to reduce the number of variables (Guisan *et al.* 1998). Our results indicate that the differences between the models constructed by PCA-based data and the original data were not significant. In PCA, each principal component reduces the remaining variance in the matrix of environmental data, and all variables contribute to all axes of PCA (Dormann *et al.* 2008). The outcome could differ for different models, e.g. Elith et al. (2011) argued that MaxEnt (a species distribution model technique) did not require PCA to avoid collinearity. Our results also demonstrated that whether PCA is required to reduce the effect of correlated predictor variables depends on the predictor variables used. Alternatively, the number of correlated predictors can be reduced before model processing.

Different models have different predictive powers (Austin 2007; Elith and Leathwick 2009; Elith *et al.* 2006; Guisan and Harrell 2000). The GLM and CART are generally

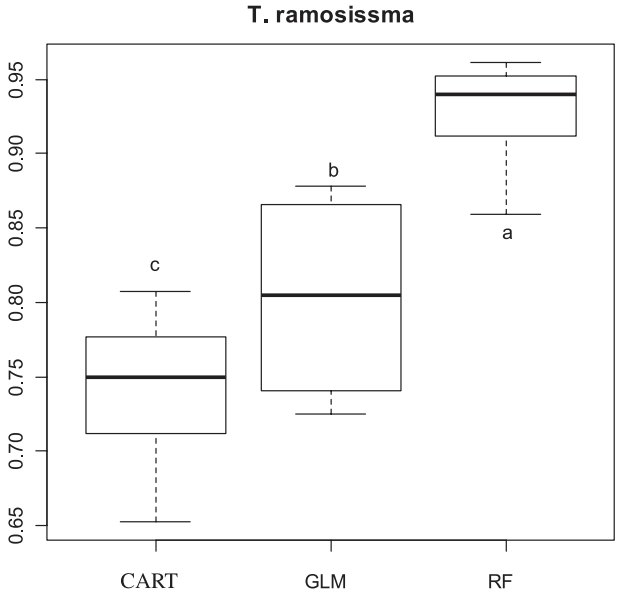

Figure 3: mean AUC for different models. The abbreviations in this figure are the same as in Table 1.

Table 1: average AUC for each *Tamarix ramosissma* model

	CART	GLM	RF
All	0.755	0.859	0.951
Cli	0.781	0.738	0.956
prcli	0.744	0.741	0.878
prall	0.690	0.874	0.929

'cli' is the dataset of climate variables, 'all' is the dataset of soil, climate and DEM variables, 'prcli' is the PCA-based dataset of climate variables 'prall' is the PCA-based dataset of soil, climate and DEM variables and 'RF' is the Random Forests model.

considered as good techniques with high prediction power (Austin 2007; Elith *et al.* 2006; Guisan and Harrell 2000; Muñoz and Felicísimo 2004). Thuiller *et al.* (2003) pointed out that classification tree analysis is less accurate than the generalized methods, especially at finer scales. The results in this study are similar to those of Thuiller *et al.* (2003). Lawler *et al.* (2006) found that random forest consistently outperformed the GLM, generalized additive models, CART, Genetic Algorithm for Rule Set Production and Artificial Neural Networks techniques. Prasad *et al.* (2006) found that Random Forests models and bagging (a tree-based model-averaging approach) consistently outperformed multivariate adaptive regression splines and regression trees in predicting the distributions of four tree species. Broennimann *et al.* (2007) modelled the distribution of *Centaurea maculosa* with BIOMOD tool and found that the performance rank, from best to worst, was Random Forests, GLM and CART. Jeschke and Strayer (2008) pointed out that new techniques, e.g. Random Forests, outperform more established methods. The results of this study indicate that the prediction precision of the Random Forests model is better than that of the GLM and CART models.

The different modelling techniques applied in this study make different assumptions about the relationships between species and their environments (Guisan and Zimmermann 2000). The choice of methods always depends on the species, dataset and question. However, the newest techniques often achieve the most accurate predictions (Jeschke and Strayer 2008). The strength of Random Forests likely lies in the power derived from averaging hundreds of different models (Breiman 2001; Lawler *et al.* 2006). In addition to providing a method for modelling complex interactions without having to specify them *a priori*, tree-based models allow the relationships between the response and the predictors to vary over the domain of the study. Therefore, we recommend using Random Forests to model species distributions because of its higher predictive power.

Different models have different assumptions to suit to different species, while different species are characterized by different environmental factors. Thus, the uncertainty and the performance of different models for different species are very complex. There are only three models, one species and three sets of environmental variables in this study, which might be insufficient to completely explain the uncertainty and the performance of species distribution models. Therefore, more models, more species and more environmental variables are still needed to the comparison work, especially at a global scale.

Because of their huge area, drylands provide a huge potential to mitigate global warming through vegetation restoration, which would increase carbon sequestration (Lal 2001, 2009). In this study, the predicted potential distribution area of *T. ramosissima* was $\sim 3.57 \times 10^4$ km^2. Annual aboveground productivity including wood and assimilation organs ranged from 1.55 to 1.74 Mg/ha (based on total ground area) or from 3.10 to 7.15 Mg/ha (in homogenous stands) for *Tamarix* vegetation (Gries *et al.* 2005). It could be inferred that the potential biomass production of *T. ramosissima* in the Tarim Basin is huge; therefore, there is great potential to mitigate global warming and produce bioenergy through restoration of *T. ramosissima* in the Tarim Basin.

CONCLUSIONS

The predictive variables for species distribution models should be chosen carefully, as the use of too many predictors might reduce the prediction power. Using PCA to reduce the correlation among predictors and enhance the accuracy of species distribution model depends on the predictor variables and the models. From the comparison of models with and without PCA-based predictors, reducing the number of correlated predictors before model processing is recommended. Among the GLM, CART and Random Forests, the best model for predicting the *T. ramosissima* distribution was Random Forests with climate variables. The soil variables considered in this study did not increase the predictive performance of the model.

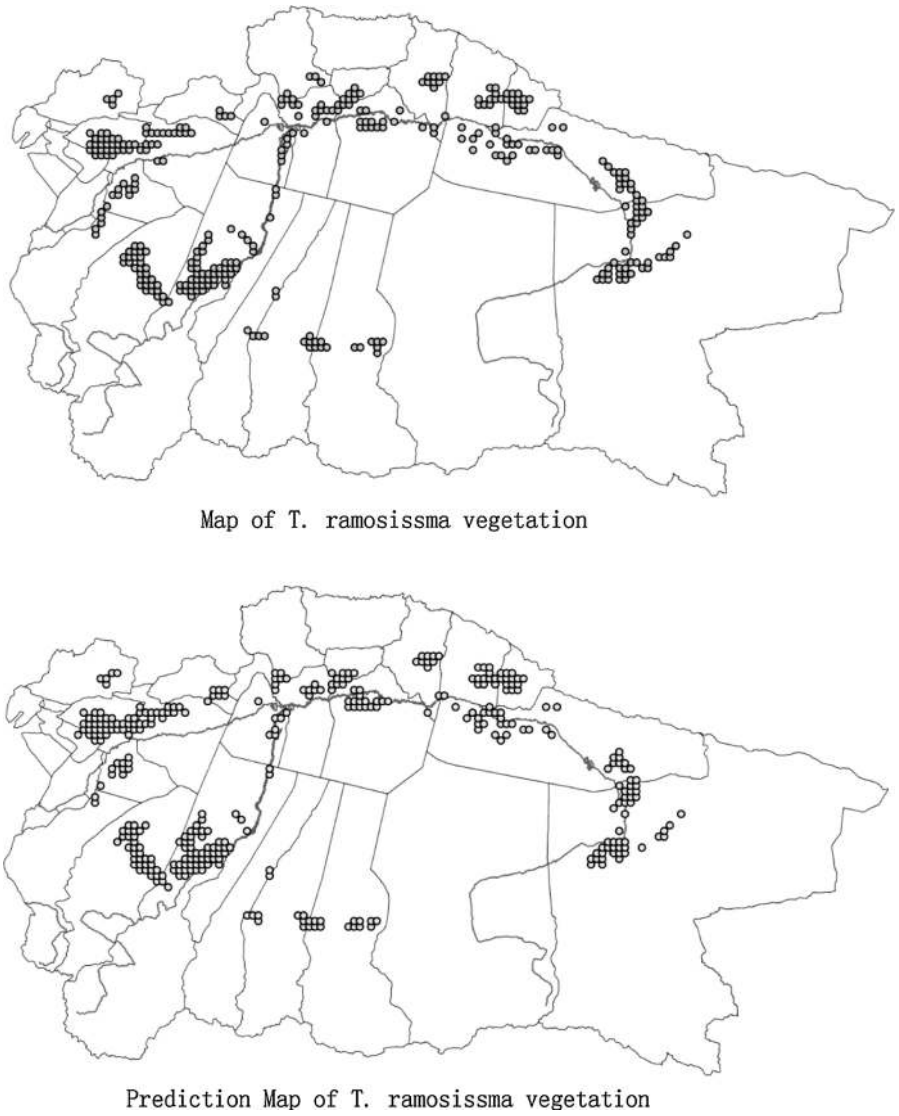

Figure 4: *Tamarix ramosissima* desert vegetation distribution and prediction map.

The Random Forests model was more precise than the GLM and CART models. The predicted potential distribution area of *T. ramosissima* was $\sim 3.57 \times 10^4$ km^2 in the Tarim Basin. In order to entirely figure out the uncertainty and the performance of different models with different species, studies with more species, more models and more data are still needed.

FUNDING

National Basic Research Program of China (973 Program) (No. 2010CB951303 and No. 2009CB421106).

ACKNOWLEDGEMENTS

The author thanks Dr Guofang Liu at Institute of Botany of CAS for processing the climate data. We would also like to thank Dr Christine Verhille at the University of British Columbia and Dr Yongbo Liu at Chinese Research Academy of Environmental Sciences for their assistance with English language and grammatical editing of the manuscript. The soil and DEM dataset were provided by the Environmental and Ecological Science Data Center for West China, National Natural Science Foundation of China.

REFERENCES

Abbott I, Le Maitre D (2010) Monitoring the impact of climate change on biodiversity: the challenge of megadiverse mediterranean climate ecosystems. *Austral Ecol* **35**:406–22.

Abideen Z, Ansari R, Khan MA (2011) Halophytes: potential source of ligno-cellulosic biomass for ethanol production. *Biomass Bioenergy* **35**:1818–22.

Allen RG, Pereira LS, Raes D, *et al.* (1998) Crop Evapotranspiration—Guidelines for Computing Crop Water Requirements. Rome, Italy: Food & Agriculture Organization of the UN.

Araujo MB, New M (2007) Ensemble forecasting of species distributions. *Trends Ecol Evol* **22**:42–7.

Austin M (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecol Modell* **200**:1–19.

Austin MP, Cunningham RB, Fleming PM (1984) New approaches to direct gradient analysis using environmental scalars and statistical curve-fitting procedures. *Plant Ecol* **55**:11–27.

Austin MP, Gaywood MJ (1994) Current problems of environmental gradients and species response curves in relation to continuum theory. *J Veg Sci* **5**:473–82.

Barry S, Elith J (2006) Error and uncertainty in habitat models. *J Appl Ecol* **43**:413–23.

Breiman L (2001) Random forests. *Mach Learn* **45**:5–32.

Breiman L, Friedman JH, Olshen RA, *et al.* (1984) Classification and Regression Trees. New York, NY: Chapman and Hall.

Broennimann O, Treier UA, Muller-Scharer H, *et al.* (2007) Evidence of climatic niche shift during biological invasion. *Ecol Lett* **10**:701–9.

Cleverly JR, Smith SD, Sala A, *et al.* (1997) Invasive capacity of Tamarix ramosissima in a Mojave Desert floodplain: the role of drought. *Oecologia* **111**:12–8.

De'ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* **81**:3178–92.

Dormann CF, Purschke O, Marquez JRG, *et al.* (2008) Components of uncertainty in species distribution analysis: a case study of the great grey shrike. *Ecology* **89**:3371–86.

Elith J, Graham CH, Anderson RP, *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29**:129–51.

Elith J, Leathwick JR (2009) Species distribution models: ecological explanation and prediction across space and time. *Annu Rev Ecol Evol Syst* **40**:677–97.

Elith J, Phillips SJ, Hastie T, *et al.* (2011) A statistical explanation of maxent for ecologists. *Diver Distributions* **17**:43–57.

Eshel A, Zilberstein A, Alekparov C, *et al.* (2010) Biomass production by desert halophytes: alleviating the pressure on food production. In: Rosen MA, Perryman R, Dodds S, et al. (ed). *Recent Advances in Energy & Environment: Proceedings of the 5th IASME/WSEAS international conference on Energy & environment (EE' 10). Stevens Point.* WI: WSEAS Press, 362–7.

Evangelista PH, Stohlgren TJ, Morisette JT, *et al.* (2009) Mapping invasive tamarisk (Tamarix) a comparison of single-scene and time-series analyses of remotely sensed data. *Remote Sens* **1**:519–33.

Feagin RA (2005) Heterogeneity versus homogeneity: a conceptual and mathematical theory in terms of scale-invariant and scale-covariant distributions. *Ecol Complex* **2**:339–56.

Feng L (2008) Halophytes promising for biomass energy resources in china. *J Biotechnol* **136**:271.

Ferrier S (2002) Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Syst Biol* **51**:331–63.

Franklin J (1998) Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. *J Veg Sci* **9**:733–48.

Garzon MB, Blazek R, Neteler M, *et al.* (2006) Predicting habitat suitability with machine learning models: the potential area of pinus sylvestris l. in the Iberian Peninsula. *Ecol Modell* **197**:383–93.

Gries D, Foetzki A, Arndt SK, *et al.* (2005) Production of perennial vegetation in an oasis-desert transition zone in NW china—allometric estimation, and assessment of flooding and use effects. *Plant Ecol* **181**:23–43.

Gries D, Zeng F, Foetzki A, *et al.* (2003) Growth and water relations of Tamarix ramosissima and populus euphratica on taklamakan desert dunes in relation to depth to a permanent water table. *Plant Cell Environ* **26**:725–36.

Guisan A, Harrell FE (2000) Ordinal response regression models in ecology. *J Veg Sci* **11**:617–26.

Guisan A, Theurillat JP, Kienast F (1998) Predicting the potential distribution of plant species in an alpine environment. *J Veg Sci* **9**:65–74.

Guisan A, Weiss SB, Weiss AD (1999) GLM versus CCA spatial modeling of plant species distribution. *Plant Ecol* **143**:107–22.

Guisan A, Zimmermann NE. (2000) Predictive habitat distribution models in ecology. *Ecol Modell* **135**:147–86.

Hamann A, Wang TL (2006) Potential effects of climate change on ecosystem and tree species distribution in British Columbia. *Ecology* **87**:2773–86.

Huisman J, Olff H, Fresco LMF (1993) A hierarchical set of models for species response analysis. *J Veg Sci* **4**:37–46.

Ibanez I, Silander JA, Wilson AM, *et al.* (2009) Multivariate forecasts of potential distributions of invasive plant species. *Ecol Appl* **19**:359–75.

Jeschke JM, Strayer DL (2008) Usefulness of bioclimatic models for studying climate change and invasive species. *Ann N Y Acad Sci* **1134**:1–24.

Jones CC, Acker SA, Halpern CB (2010) Combining local- and large-scale models to predict the distributions of invasive plant species. *Ecol Appl* **20**:311–26.

Lal R (2001) Potential of desertification control to sequester carbon and mitigate the greenhouse effect. *Clim Change* **51**:35–72.

Lal R (2009) Sequestering carbon in soils of arid ecosystem. *Land Degrad Dev* **20**:441–54.

Larssen T, Hogasen T, Cosby BJ (2007) Impact of time series data on calibration and prediction uncertainty for a deterministic hydrogeochemical model. *Ecol Modell* **207**:22–33.

Lawler JJ, White D, Neilson RP, *et al.* (2006) Predicting climate-induced range shifts: model differences and model reliability. *Global Change Biol* **12**:1568–84.

Li X, Huang Y, Gong J, *et al.* (2009) A study of the development of bio-energy resources and the status of eco-society in china. *Energy* **35**:4451–6.

Liu MT (1995) Synthesis Study and Expanding Application for Plants from Genus of Tamarix L. *[in Chinese]*. Lanzhou, China: Lanzhou University Press.

McCullagh P, Nelder JA. (1989) Generalized Linear Models. London, UK: Chapman & Hall/CRC.

Mckenney DW, Pedlar JH, Lawrence K, *et al.* (2007) Potential impacts of climate change on the distribution of North American trees. *Bioscience* **57**:939–48.

Mellin C, Bradshaw CJA, Meekan MG, *et al.* (2010) Environmental and spatial predictors of species richness and abundance in coral reef fishes. *Global Ecol Biogeogr* **19**:212–22.

Muñoz J, Felicísimo ÁM (2004) Comparison of statistical methods commonly used in predictive modelling. *J Veg Sci* **15**:285–92.

Pearson RG, Dawson TP (2003) Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecol Biogeogr* **12**:361–71.

Peters J, Verhoest NEC, Samson R, *et al.* (2009) Uncertainty propagation in vegetation distribution models based on ensemble classifiers. *Ecol Modell* **220**:791–804.

Phillips DL, Marks DG (1996) Spatial uncertainty analysis: propagation of interpolation errors in spatially distributed models. *Ecol Modell* **91**:213–29.

Prasad AM, Iverson LR, Liaw A (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* **9**:181–99.

R Development Core Team (2010) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Randin CF, Engler R, Normand S, *et al.* (2009) Climate change and plant distribution: local models predict high-elevation persistence. *Global Change Biol* **15**:1557–69.

Ray N, Burgman MA (2006) Subjective uncertainties in habitat suitability maps. *Ecol Modell* **195**:172–86.

Retuerto R, Carballeira A (2004) Estimating plant responses to climate by direct gradient analysis and geographic distribution analysis. *Plant Ecol* **170**:185–202.

Rotenberry JT, Preston KL, Knick ST (2006) Gis-based niche modeling for mapping species' habitat. *Ecology* **87**:1458–64.

Stromberg JC, Lite SJ, Marler R, *et al.* (2007) Altered stream-flow regimes and invasive plant species: the Tamarix case. *Global Ecol Biogeogr* **16**:381–93.

Tang Y, Xie J-S, Geng S (2010) Marginal land-based biomass energy production in china. *J Integr Plant Biol* **52**:112–21.

Thuiller W (2003) Biomod—optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biol* **9**:1353–62.

Thuiller W, Araujo MB, Lavorel S (2003) Generalized models vs. classification tree analysis: predicting spatial distributions of plant species at different scales. *J Veg Sci* **14**:669–80.

Thuiller W, Lafourcade B, Engler R, *et al.* (2009) Biomod—a platform for ensemble forecasting of species distributions. *Ecography* **32**:369–73.

Townsend PA, Pape M, Eaton M (2007) Transferability and model evaluation in ecological niche modeling: a comparison of garp and maxent. *Ecography* **30**:550–60.

van Horssen PW, Pebesma EJ, Schot PP (2002) Uncertainties in spatially aggregated predictions from a logistic regression model. *Ecol Modell* **154**:93–101.

Van Niel KP, Austin MP (2007) Predictive vegetation modeling for conservation: impact of error propagation from digital elevation data. *Ecol Appl* **17**:266–80.

Yang WK, Yin LK, Zhang DY, *et al.* (2004) Study on ecological types and habitat similarity of Tamarix L. in Xinjiang [in Chinese]. *Arid Land Geogr* **27**:186–92.

Yates CJ, Elith J, Latimer AM, *et al.* (2010) Projecting climate change impacts on species distributions in megadiverse South African Cape and Southwest Australian Floristic Regions: opportunities and challenges. *Austral Ecol* **35**:374–91.

Yee TW, Mitchell ND (1991) Generalized additive-models in plant ecology. *J Veg Sci* **2**:587–602.

Zhang Q (2011) Simulation the potential geographical distribution and evaluation the restoration of Tamarix vegetation in Tarim Basin [in Chinese]. *Ph.D. Thesis*. Graduate University of Chinese Academy of Sciences. Beijing.

Zhang XM, Runge M (2006) Ecological Basis for Sustainable Managing the Vegetation in the Fringe of Taklimakan Desert *[in Chinese]*. Beijing, China: Science Press.

Zhang XS (2008) China 1:100 Million Vegetation Map *[in Chinese]*. Beijing, China: Geological Publishing House of China.