Remote Sensing Time Series for Modeling Invasive Species Distribution: A Case Study of *Tamarix* spp. in the US and Mexico

Anna Cord\(^{a,b}\), Doris Klein\(^b\), Stefan Dech\(^{a,b}\)

\(^a\) German Remote Sensing Data Center, German Aerospace Center (DLR-DFD), Münchner Str. 20, 82234 Wessling, Germany (anna.cord@dlr.de, stefan.dech@dlr.de)

\(^b\) Department of Remote Sensing, Institute of Geography, University of Wuerzburg, Am Hubland, 97074 Wuerzburg, Germany (anna.cord@uni-wuerzburg.de, doris.klein@uni-wuerzburg.de, stefan.dech@uni-wuerzburg.de)

**Abstract:** Detecting invasive species and predicting their potential distribution are crucial to coordinate management responses. Remote sensing data are now available in several spatial and temporal resolutions and can supply environmental models with additional information. This study uses the Maximum Entropy algorithm to model the current distribution of the saltcedar (*Tamarix* spp.) in the US and Mexico and to identify suitable habitats, both already inhabited and not yet occupied. Tamarisk is restricted to specific habitats such as riparian zones, wetlands and agricultural or disturbed areas, which are typically not only characterized by climate. To describe vegetation phenology and thermal seasonality in these habitats, the study uses annual metrics of remotely sensed time series from 2001 to 2008 (Terra-MODIS Enhanced Vegetation Index and Land Surface Temperature) together with WorldClim bioclimatic data. By using occurrence records primarily from the US we were able to model predictive maps of tamarisk distribution correlating very well to the known distribution in the US. For Mexico, where only very few occurrence records exist, we identified potential tamarisk habitats for substantial areas in Baja California, in the states of Sonora and Sinaloa and in the Central Mexican Plateau. These predictive model results can be used to support the early detection and prevention of *Tamarix* spp. invasion.

**Keywords:** Maximum Entropy; MODIS Time Series; Enhanced Vegetation Index; Land Surface Temperature; Tamarisk.

**Remark on the nomenclature of this study:** *Tamarix* spp. refers to all species within the genus *Tamarix* L., primarily *T. ramosissima* and *T. chinensis* and hybrids.

1. **INTRODUCTION**

On a global scale, invasive species belong to the main reasons of biodiversity loss [Ficetola et al. 2007] and cause substantial economic damage. A large number of studies are therefore investigating species invasions with the aim to identify suitable habitats and to coordinate management plans. The conceptual findings and spatio-temporal predictions of these studies contribute to the early detection and elimination of invasive species to reduce their negative impacts. The saltcedar (*Tamarix* spp.) is native to Southern Europe, Asia Minor, Mongolia, Tibet, central China, and North Korea [Carpenter 2003] and has the growth habit of a shrub or small tree. In the 1800s, it was introduced in the United States as ornamental plant and as windbreak for erosion control [Pearce and Smith 2003]. Tamarisk is a facultative phreatophyte, well-adapted to arid environments and – due to its extensive
root system – better able to tolerate drought, water stress, and salinity compared to native mesic species [Glenn and Nagler 2005; Kerns et al. 2009]. Since its introduction, *Tamarix* spp. has become the (sub)dominant woody species along many perennial river systems in the arid south-western United States and north-western Mexico and has increasingly been replacing native vegetation [Glenn and Nagler 2005]. By the 1920s, tamarisk was first recognized as a problem, having several detrimental effects such as decreased stream flow, loss of native biodiversity, changes in food web structure and soil salinization [DiThomaso 1998]. The decline of many native riparian phreatophytes connected with *Tamarix* spp. expansion has been related to altered river flow regimes due to water management programs and the use of ground water for urban, agricultural and industrial purposes [Shafroth et al. 1998]. Today, the tamarisk is considered among the “World’s Worst Invasive Alien Species” [Lowe 2000] and one of the most aggressively invasive exotic plants in the US and Mexico. Once the species is dominant at a site, control and restoration efforts are extremely costly in terms of labor, time and money. Zavaleta [2000] calculated the annual costs resulting from tamarisk invasion to be between US$ 280 and 450/ha with almost one million ha already covered in the western United States alone [Pearce and Smith 2003]. Due to its high invasion potential, several previous studies aimed at mapping tamarisk distribution on small scales – often using high spatial resolution [Ge et al. 2006] or hyperspectral data [Hamada et al. 2007]. In addition, previous work [Everitt and DeLoach 1990] showing the specific temporal-spectral signature of *Tamarix* spp. related to vegetation phenology suggests the high potential of remotely sensed time series data in this context. Morissette et al. [2006] used land cover data and three phenological metrics derived from four years of MODIS vegetation index data in a logistic regression approach to create a habitat suitability map for the continental US. Recently, Evangelista et al. [2009] showed that multi-temporal Landsat imagery significantly improved the prediction of tamarisk occurrence compared to single-scene analysis for a small study area in Colorado. The current center of tamarisk distribution includes the states of Arizona, New Mexico and Utah [Glenn and Nagler 2005]. Its distribution in Mexico is not well-known but already documented for the Sonoran coast [Harrison and Matson 2003], the Colorado River delta [Glenn et al. 1996] and Guaymas (Sonora, Mexico) [West and Nabhan 2002]. Building on the findings of previous studies such as Morissette et al. [2006], this study uses multi-year (2001-2008) time series of remotely sensed data to model the distribution of *Tamarix* spp. throughout the southern US and Mexico. We analyzed the Terra-MODIS Vegetation Index and Land Surface Temperature products regarding their usefulness for predicting potential tamarisk distribution together with *Tamarix* spp. occurrence records using the Maximum Entropy (Maxent) algorithm [Phillips et al. 2006]. Apart from Morissette et al. [2006], we are not aware of any other study modeling *Tamarix* spp. habitat covering comparatively large areas with remote sensing time series data in combination with topo-climatic data. Another novel aspect of this study is the approach to reliably predict the as of yet unknown potential range of tamarisk in Mexico.

2. DATA AND METHODS

2.1. *Tamarix* spp. records

We collected *Tamarix* spp. occurrence records from the Global Biodiversity Information Facility (GBIF, http://www.gbif.org) and from the National Institute of Invasive Species Science (NIISS, http://www.niiss.org) covering the current center of tamarisk distribution in the US in California, Arizona, Utah, New Mexico, Colorado and Texas (see Figure 2d). While GBIF is a global data base primarily including herbarium and museum collections, NIISS compiles occurrence records of invasive species for the US alone. NIISS records were mainly collected during intensive species-specific field mapping projects such as – for tamarisks – the Southwest Exotic Mapping Program (SWEMP), Arkansas River Watershed Invasive Plant Plan (ARKWIPP) or helicopter surveys, and thus covered a wider range of habitats compared to GBIF samples. Due to a lack of sampling effort, only few samples were available for known populations in northern Mexico. Since model quality and results can be significantly influenced when samples are not representative for the whole climatic niche space of the target species, we used the Mexican *Tamarix* spp. records (10 samples) to run separate models. We are aware that this sample size is at the
lower limit to produce stable model predictions. Another important concern in species records is spatial autocorrelation, which may significantly impact model predictions [Dormann et al. 2007]. Spatial autocorrelation is basically a lack of independence between observations due to the fact that vicinity in space impacts the chance of occurrence. To reduce the effects of inherent spatial autocorrelation (especially between training and test sub-samples), we removed all *Tamarix* spp. records within the same 10 arc-minutes grid cells. The resulting data base comprised 1,726 records (NIISS: 1,449 and GBIF: 277).

### 2.2. Pseudo-absence / Background points

As recent work has shown [Philipps 2009], the influence of spatially biased samples can be reduced by comparing the occurrences with background points which reflect the same spatial bias (rather than using random background points). The idea is that a model based on presence and background data with the same bias will not focus on the sample selection bias but on any differentiation between the distribution of the occurrences and that of the background. Background data should thus be collected using the same methods or equipment as those used for the target species data. We therefore used 1,944 GBIF records of *Salix* spp. (willow), native trees often replaced during tamarisk invasion [Glenn and Nagler 2005]. Both *Tamarix* and *Salix* records showed the same spatial pattern with a concentration along rivers / stream lines and a core area of distribution in the southern US.

### 2.3. Remote sensing data

Two global MODIS-Terra L3 standard products of nominal 1x1 km² spatial resolution were used: (1) *Vegetation Indices* (MOD13A2, 16-day composites, Collection 5) designed for vegetation studies and (2) *Land Surface Temperature* (LST, MOD11A2, 8-day composites, Collection 5). The MODIS data were acquired from January 2001 through December 2007 (MOD13A2) / December 2008 (MOD11A2), as available through the *NASA Earth Observing System Data Gateway* (https://wist.echo.nasa.gov/api/). Nine MODIS tiles were mosaicked and reprojected to geographic projection (Datum WGS 1984) using freely available *MODIS Reprojection Tool* software (MRT, Version 4). To identify and replace low quality data, the *Time Series Generator* (TiSeG) software [Colditz et al. 2008] for linear temporal interpolation between valid observations was applied. Annual metrics (Table 1) as measures of vegetation greenness, phenology and thermal seasonality were calculated from the time series and its first derivative and averaged over the seven / eight years of the study period. We decided to use these metrics as the presence or absence of a species in any area is often distinguished not only by absolute levels of vegetation or climate variables, but also by subtle differences in their seasonality.

### 2.4. Bioclimatic and topographic data

The climate data set used (WorldClim, Version 1.4, http://www.worldclim.org) is based on temperature and precipitation values in the period 1950 to 2000 of a global network of 4,000 climate stations [Hijmans et al. 2005]. The meteorological station data had been interpolated to monthly climatic surfaces by using a thin-plate smoothing spline algorithm with latitude, longitude, and elevation data as independent variables [Hijmans et al. 2005]. The full set of 19 bioclimatic variables derived from this WorldClim data set was downloaded at 30 arc-seconds resolution, resampled and gridded to pixel location and cell size of the MODIS data. These bioclimatic parameters express spatial variations in annual means, seasonality and extreme or potentially limiting climatic factors for ecological studies. Since *Tamarix* spp. primarily occurs in riparian habitats [Glenn and Nagler 2005], we also calculated slope values (percent rise) from the SRTM digital elevation model available at an aggregated resolution of 30 arc-seconds through the WorldClim webpage. All environmental data were prepared and converted to ASCII files using DIVA-GIS software (Version 5.4, http://www.diva-gis.org).

### 2.5. Selection of environmental predictors

To improve computing efficiency, we performed a correlation analysis between all environmental variables (19 bioclimatic WorldClim variables, slope, 14 vegetation index...
metrics and six land surface temperature metrics). We retained only one of several correlating variables with a Pearson correlation coefficient of $r^2 > 0.75$ from the available input data to act as a proxy for the other co-variables with respect to the eco-physiological requirements of tamarisk as described by Glenn and Nagler [2005].

2.6. Maximum Entropy model

We conducted our analysis using the Maxent modeling software (Version 3.3.1, available from www.cs.princeton.edu/~schapire/maxent/), which has great potential for identifying species distributions and habitat selection patterns [Baldwin 2009] and proved to be very useful in several comparative studies compared to other algorithms such as BIOCLIM, GAM, DOMAIN or GARP [Elith et al. 2006]. Maxent is a general-purpose algorithm for estimating probability distributions based on the principle of Maximum Entropy [Phillips et al. 2006]. The model evaluates the continuous suitability of each grid cell as a function of environmental variables from 0 (unsuitable habitat) to 1 (optimal habitat). Some advantages of Maxent are that it requires presence-only data, appears to be less sensitive than other approaches to the number of presence locations, can incorporate interactions between different variables (by automatically computing “product features” between predictor variables) and has no specific underlying assumptions regarding the statistical properties of the input data [Phillips et al. 2006]. However, one drawback of the Maxent method is that it uses an exponential model for probabilities which is not inherently bounded above. Thus, very high predicted values for environmental conditions outside the range present in the study area may occur [Phillips et al. 2006]. Maxent models were run under the following settings: 10 replicates (bootstrapping), auto features, jackknife test = true, logistic output format (ASCII), random test percentage = 25, regularization multiplier = 1, maximum iterations = 500, convergence threshold = 0.0001, and maximum number of background points = 5,000.

3. RESULTS AND DISCUSSION

3.1. Environmental predictors

Based on the results of the correlation analysis we selected three sets of environmental predictors (see Table 1): (1) TOPOCLIM: 10 topo-climatic variables, (2) RS: 10 remote sensing variables and (3) TOPOCLIM_RS: each top 5 variables of topo-climatic and remote sensing data in sets (1) and (2) measured by variable importance. During each step of the algorithm, Maxent increases the gain of the model by modifying the coefficient for a single feature which is then assigned to the environmental variable(s) that the feature depends on [Phillips et al. 2006]. According to these results, the most important topo-climatic predictor variables were precipitation of the warmest quarter, isothermality and temperature seasonality (Table 1). Out of the remote sensing-derived variables, minimum and maximum surface temperatures together with annual surface temperature range turned out to be the major determinants (Table 1) and were significantly more important than vegetation index features. The climatic parameters observed at tamarisk presence sites are characteristic of continental warm and arid climates. Our results are thus consistent with other studies, e.g. Kerns et al. [2005], who found that Tamarix spp. occurrence is positively correlated to warmer and drier sites with maximum daily surface temperature being the most important variable in an ENFA model. According to Friedman et al. [2005], the frequency of occurrence of T. ramosissima also has a strong positive relation with the mean annual minimum temperature (due to frost sensitivity). This hypothesis was confirmed by the high variable importance of the annual minimum surface temperature in our Maxent model. We assume that the spatial detail of BIO6 (minimum temperature of the coldest month) is not sufficient to reproduce the same relationship within the topo-climatic predictors. For the successful prediction of species invasions between currently occupied and target regions, environmental similarity is the basic requirement [Ficetola et al. 2007]. Several of our predictor variables show a strong north-south gradient, so that the spatial bias of the species occurrence records translates into a bias in environmental predictor space between records in the northern (US) and southern (Mexico) part of the study region.
A. Cord et al. / Remote sensing data for monitoring and modeling invasive species distribution

(Figure 1). This effect was taken into account for the distribution model (Sections 3.2. and 3.3.).

Table 1. Environmental predictors of the TOPOCLIM and RS data sets – Data source, ecological / bio-physiological interpretation and variable importance.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Information</th>
<th>Importance (%)</th>
<th>Variable</th>
<th>Information</th>
<th>Importance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIO18</td>
<td>Precipitation of warmest quarter of the year</td>
<td>32.32</td>
<td>LSTmin</td>
<td>Minimum surface temperature</td>
<td>25.38</td>
</tr>
<tr>
<td>BIO3</td>
<td>Isothermality</td>
<td>19.40</td>
<td>LSTmax</td>
<td>Maximum surface temperature</td>
<td>21.40</td>
</tr>
<tr>
<td>BIO4</td>
<td>Temperature seasonality (standard deviation * 100)</td>
<td>16.69</td>
<td>LSTrange</td>
<td>Annual surface temperature range</td>
<td>13.97</td>
</tr>
<tr>
<td>BIO1</td>
<td>Annual mean temperature</td>
<td>11.04</td>
<td>LSTmean</td>
<td>Mean surface temperature</td>
<td>8.47</td>
</tr>
<tr>
<td>Slope</td>
<td>Terrain steepness</td>
<td>8.21</td>
<td>EVImax</td>
<td>Maximum green biomass</td>
<td>7.17</td>
</tr>
<tr>
<td>BIO6</td>
<td>Min. Temperature of coldest month</td>
<td>4.50</td>
<td>EVIdatenum</td>
<td>Time of peak of growing season</td>
<td>7.08</td>
</tr>
<tr>
<td>BIO15</td>
<td>Precipitation seasonality</td>
<td>2.07</td>
<td>EVImin</td>
<td>Minimum green biomass</td>
<td>5.56</td>
</tr>
<tr>
<td>BIO14</td>
<td>Precipitation of driest month</td>
<td>2.04</td>
<td>EVIdatemin</td>
<td>Time of minimum vegetation greenness</td>
<td>5.31</td>
</tr>
<tr>
<td>BIO2</td>
<td>Mean diurnal temperature range</td>
<td>1.90</td>
<td>EVIdevdate max</td>
<td>Date of vegetation green-up</td>
<td>4.02</td>
</tr>
<tr>
<td>BIO19</td>
<td>Precipitation of coldest quarter of the year</td>
<td>1.84</td>
<td>EVIdevdate min</td>
<td>Date of vegetation senescence</td>
<td>1.64</td>
</tr>
<tr>
<td>SUM</td>
<td></td>
<td>100.00</td>
<td>SUM</td>
<td></td>
<td>100.00</td>
</tr>
</tbody>
</table>

Figure 1. Comparison of mean Land Surface Temperature (LST) values for *Tamarix* spp. records in the US and Mexico compared to 10,000 random background points. The background points from USA and Mexico were significantly different (Mann-Whitney U test, p<0.001).

3.2. Model performance

ROC (Receiver Operating Characteristic) plots were obtained by plotting all sensitivity values (true positive fraction) on the y-axis against their equivalent values (1-specificity, false positive fraction) on the x-axis. Maxent treats the background pixels as negative instances and species presence pixels as positive instances. The resultant AUC (Area Under Curve) is scaled between 0.0 and 1.0 and provides a measure of model accuracy that is independent of a particular probability cut-off and has therefore become one of the standard methods for assessing performance of SDMs [Anderson et al. 2003]. We obtained: AUC\textsubscript{TOPOCLIM} = 0.70 (prevalence 0.38), AUC\textsubscript{RS} = 0.68 (prevalence 0.41) and AUC\textsubscript{TOPOCLIM, RS} = 0.71 (prevalence 0.39). According to a common classification [Swets 1988], 0.9 describes ‘very good’, 0.8 ‘good’ and 0.7 ‘useable’ discrimination ability of the respective model. The – according to these standards – ‘poor’ performance of our models is due to the application of *Salix* spp. occurrence records as background data, which leads to *Tamarix* spp. being modeled as more of a generalist than a specialist. Generalist species are not as well discriminated from the background as specialists (whose occurrences are more spatially and environmentally clustered) and thus result in lower AUC values [Veloz 2009]. Analogous Maxent models run with random background points covering the entire study
area revealed AUC values between 0.84 and 0.91 (prevalence 0.22 – 0.28). Nevertheless, we decided rather to correct for the prominent sampling bias present in the species records than to go for the maximum AUC. Using the reduced set of only Mexican occurrence records (10 samples), we obtained significantly different results ($AUC_{\text{TOPOCLIM}} = 0.98$ (prevalence 0.05), $AUC_{R} = 0.96$ (prevalence 0.09) and $AUC_{\text{TOPOCLIM}_R} = 0.99$, prevalence 0.05). The combination of TOPOCLIM and RS data resulted in all cases in the highest AUC values. Absence points (even though available through NIIS) were not useful for model validation because tamarisk is currently not at its maximum potential distribution.

3.3. Modeled *Tamarix* spp. distribution

Applying the different sets of environmental variables resulted in predictive maps (Figure 2) well-related to the known distribution [e.g. Glenn and Nagler 2005, Morisette et al. 2006] of *Tamarix* spp. when using all occurrence records (Figure 2 (a) – (c)). However, the model predictions suggest that there is more suitable habitat than is currently invaded and indicate that we may still be in the early stage of tamarisk invasion. Morisette et al. [2006] found that especially low- and mid-elevation waterways in the western and central US are still susceptible to tamarisk invasion.

![Figure 2](image-url)

*Figure 2.* Modeled potential tamarisk distribution using the TOPOCLIM (a), RS (b) and TOPOCLIM_Rs (c) sets of environmental variables for all records and (d) – (f) only Mexican records respectively. The distribution of GBIF / NIIS records is shown in (d).

For Mexico, potential tamarisk habitat was mapped for large areas in Baja California, in the states of Sonora and Sinaloa and in the Central Mexican Plateau (Figure 2 (d) – (f)).
According to these model results, we assume a high probability of tamarisk invasion into (semi)arid regions in Mexico in the future. Climate change, hybridization and continued human impact on river flow regimes may even facilitate this trend.

Some caveats of the maps should be mentioned: First, pathways or sources for the introduction of Tamarix spp. were not available – even though these data are essential prerequisites for the establishment of invasive species. Secondly, the distribution maps are produced at a coarse 1 km² resolution constrained by the amount and individual file size of model input data. Therefore, they show not the actual presence of tamarisk but rather the probability of occurrence or habitat suitability (measured by similarity to vegetation and habitat characteristics of hot spots of tamarisk occurrence). Small-scale habitat patches (such as springs, narrow riparian zones, small tributaries) might thus not be detected by the model. It is therefore appropriate to use our results for large-scale assessments or to identify focus areas to be analyzed with higher resolution environmental data or remote sensing imagery.

4. CONCLUSION

Field surveys, remote sensing data and species distribution models are important tools for identifying priority areas for rapid response and developing efficient, long-term control strategies over large areas for any invasive species. In our opinion, the best interpretation of such model predictions is not as an absolute map of probability of occurrence but rather as a relative ranking of suitable habitat conditions since the true proportion of habitat already occupied by the target species is still difficult to assess by predictive modeling. The Maxent algorithm proved to be very useful in this study and has several features that support the analysis of remote sensing data. The different model scenarios showed that remote sensing time series contributed significantly to discover habitat characteristics even within similar climatic conditions. Even though the environmental data used for this study do not fully cover the spatial extent of currently known Tamarix spp. distribution, the models successfully reproduced its known distribution within the southern US. Additionally, we were able to predict the distribution for Mexico, where very little occurrence data exist, but where tamarisk is already recognized as an invasive species. To our knowledge, this distribution map is the first approximation towards modeling the potential range in Mexico. Model predictions produced in this study show that there is a very high risk of Tamarix spp. invasion over large areas in Mexico. Comprehensive tamarisk field surveys are thus required to confirm the preliminary results of this study. Future work will also have to include the analysis of higher resolution remotely sensed time series.

REFERENCES


Ge, A., Carruthers, R., Gong, P. and A. Herrera, Texture analysis for mapping Tamarix parviflora using aerial photographs along the cache creek, California, Environmental Monitoring and Assessment, 114, 65-83, 2006.


