International Environmental Modelling and Software Society (iEMSs)
2012 International Congress on Environmental Modelling and Software
Managing Resources of a Limited Planet, Sixth Biennial Meeting, Leipzig, Germany
R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.)
http://www.iemss.org/society/index.php/iemss-2012-proceedings

Suitability of land cover and remote sensing data for modelling species distributions

Anna F. Cord[a], Doris Klein[b], Franz Mora[c], Stefan Dech[b,d]
[a] *Helmholtz Centre for Environmental Research (UFZ), Permoserstr. 15, 04318 Leipzig*
[b] *German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), Münchner Straße 20, 82234 Weßling, Germany*
[c] *National Commission for the Knowledge and Use of Biodiversity (CONABIO), Avenida Liga Periférico-Insurgentes Sur 4903, Col. Parques del Pedregal, Del. Tlapan, 14010, Mexico-City, Mexico*
[d] *Department of Remote Sensing, University of Würzburg, Am Hubland, 97074 Würzburg, Germany*
anna.cord@ufz.de

Abstract: Current changes of biodiversity result almost exclusively from human activities. As a consequence, spatially continuous estimates of species distributions are needed to support biodiversity evaluation and management. In the last two decades, *species distribution models* (SDMs) have been established as important tools for extrapolating *in situ* (point) observations. To account for current habitat loss, climate data used as predictors in SDMs need to be complemented by measures of current land surface characteristics. For this purpose, two alternative data sources are available, namely categorical land cover and continuous remote sensing data, each with their advantages and drawbacks. The objective of this study was therefore to directly compare the suitability of an existing land cover classification and remote sensing time series for the delineation of current biotope availability. The analysis used the *Maximum Entropy* algorithm to model the distributions of twelve tree species representative of the major Mexican forest types. Model results were evaluated based on AUC (*area under curve*) and statistical model deviance and revealed that land cover-based models overestimated species distributions and that the suitability of land cover data was dependent on species characteristics. The findings of this study support the selection of predictors in species distribution modelling in the future.

***Keywords**:* Remote sensing; Land cover; Species distribution model; Mexico.

1 INTRODUCTION

Spatial decision support systems for biodiversity evaluation and management rely on spatially continuous rather than point species data. Therefore, understanding and monitoring species distributions is crucial for nature conservation management. The most effective way to maximize the information content on species locality data is to apply *species distribution models* (SDMs) based on environmental characteristics. Habitat information often used in SDMs to refine climatic distribution ranges can be indirectly obtained from land cover maps derived from remote sensing observations. There are currently a number of continental or global mapping activities ongoing and many land cover products are freely distributed, e.g. IGBP DISCover, GLC 2000 or GLOBCOVER. The discrete representation of land surface characteristics in these products “has the advantages of concision and clarity" and "represents low data volume" (Lambin, 1999: p. 193). As land cover data are delivered in 'ready-to-use' raster formats including the required metadata information, ecologists increasingly integrate such data into their models. However, land cover maps are often not (thematically) detailed enough (Bradley and Fleishman, 2008), which supports the present trend towards the direct integration of continuous spectral remote sensing data or derived remote sensing vegetation indices into SDMs. Nevertheless, using remote sensing data as primary data for modelling purposes requires the analysis of high data volumes. The question therefore arises whether the use of such remote sensing data is worth the pre-processing effort compared to readily available land cover data. Both data sources further have a different measurement scales, namely categorical or continuous, as well as certain advantages and drawbacks. Their usefulness is therefore subject to an active scientific discussion. The objective of this study is to assess and compare the suitability of multi-temporal remote sensing data and an existing categorical land cover classification for modelling tree species distributions in Mexico.

2 STUDY AREA AND SPECIES

Mexico exhibits a great environmental and biological diversity that is reflected in an enormous variety of ecological processes and high levels of species richness and endemism (Sarukhán et al., 2010). In line with the global trend, the greatest threat to biodiversity in Mexico is the loss of habitats, especially the deforestation of natural ecosystems for food production (Sarukhán et al., 2010). The twelve study tree species are representative of the major Mexican forest types and were chosen to capture a wide variety of ecological traits such as range size or biotope specificity (Table 1). Forests belong to the vegetation types in Mexico with the highest species numbers and are particularly threatened by extensive transformation for agriculture or infrastructural activities (Ricker et al., 2007).

Table 1. Overview of the study species including range size, typical vegetation type, and number or presence records.

Species	Range size	Vegetation type	Records
Abies religiosa	restricted	Temperate needle leaved evergreen forest	132
Alnus acuminata	wide	Temperate deciduous-evergreen forest	229
Arbutus xalapensis	wide	Temperate deciduous-evergreen forest	2,530
Astronium graveolens	wide	Tropical or sub-tropical evergreen forest	438
Avicennia germinans	wide	Wetlands	54
Bursera bipinnata	intermediate	Tropical or sub-tropical deciduous forest	241
Bursera simaruba	wide	Tropical or sub-tropical evergreen forest	3,940
Cedrela odorata	wide	Tropical or sub-tropical evergreen forest	381
Guaiacum sanctum	restricted	Tropical or sub-tropical deciduous-evergreen forest	118
Liquidambar macrophylla	restricted	Moist montane/cloud forest	91
Liquidambar styraciflua	intermediate	Moist montane/cloud forest	127
Pinus chiapensis	restricted	Temperate needle leaved evergreen forest	21

3 DATA AND METHODS

3.1 Species occurrence data

Species occurrence data used stem from the Mexican *National Forest Inventory*, which was carried out by the *National Forestry Commission* (CONAFOR, *Comisión*

Nacional Forestal) between October 2004 and November 2007 to monitor a total of 24,659 sites. Distances between INFyS sites range from 5 km (for forests), through 10 km (dry forests, mangroves, wetlands) to 20 km (matorral). The data set is hence spatially biased towards forested sites. The reference area of each INFyS site is 1 ha. Given the spatial resolution of the environmental predictors used in this study, sites where a certain target species was not found, were not judged as true absences since non-detection of the species in the 1 ha reference area does not necessarily imply species absence within the corresponding 1 km^2. Instead, to account for the spatial sampling bias towards forested areas inherent in the INFyS data set, non-recorded presence was treated in the sense of the *target-group background* approach (Anderson et al., 2003) which makes use of background data that were collected with the same spatial bias as the presence records. This approach has already successfully been applied to Maxent (Mateo et al., 2010).

3.2 Land cover classification

The land cover information used for this analysis had been produced in the context of the *North American Land Change Monitoring System* project (NALCMS, 2005; Figure 1). The overall accuracy of the NALCMS land cover product was estimated at 82% for Mexico (Colditz et al., 2010). It is hence the most accurate data set currently available for this area; highest mapping accuracies were generally ascertained for forest classes (relevant for this study), lowest accuracies for barren land and temperate shrubland (Colditz et al., 2010). In addition, the land cover data are temporally corresponding to the species occurrence and remote sensing data, which is a required assumption for reliable species distribution modelling.

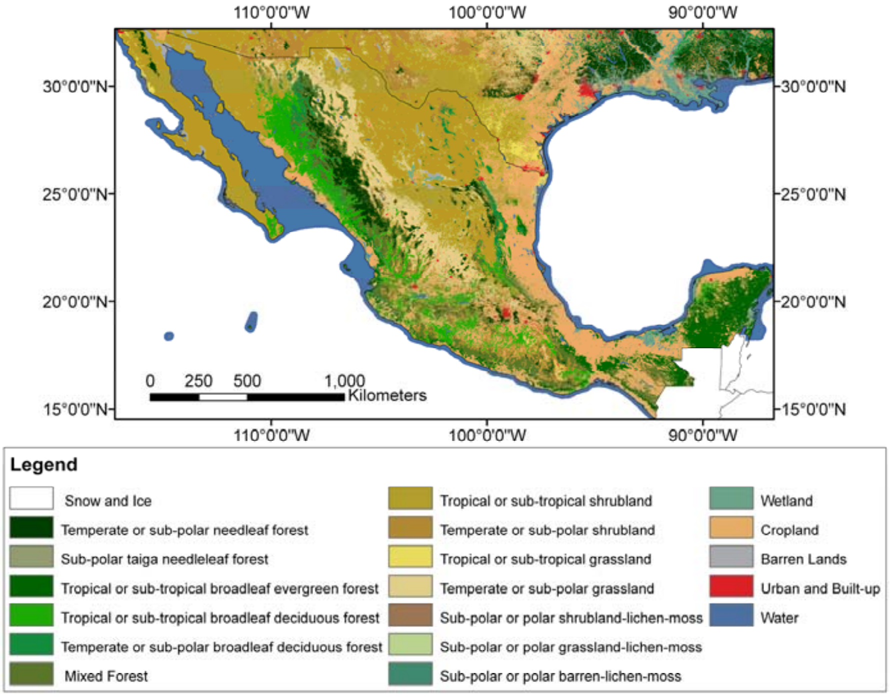
110°0'0"W
100°0'0"W
90°0'0"W
30°0'0"N
25°0'0"N
20°0'0"N
15°0'0"N
0
250
500
1,000
Kilometers
Legend
Snow and Ice
Temperate or sub-polar needleaf forest
Sub-polar taiga needleleaf forest
Tropical or sub-tropical broadleaf evergreen forest
Tropical or sub-tropical broadleaf deciduous forest
Temperate or sub-polar broadleaf deciduous forest
Mixed Forest
Tropical or sub-tropical shrubland
Temperate or sub-polar shrubland
Tropical or sub-tropical grassland
Temperate or sub-polar grassland
Sub-polar or polar shrubland-lichen-moss
Sub-polar or polar grassland-lichen-moss
Sub-polar or polar barren-lichen-moss
Wetland
Cropland
Barren Lands
Urban and Built-up
Water

Figure 1. Land cover data used for modelling species distributions. Data source: *North American Land Change Monitoring System* project (NALCMS, 2005).

3.3 Remote sensing data

In this study, time series of two MODIS 16-day standard products (1 km, Collection 5) over the 9-year period from January 2001 to December 2009 were produced. Nine MODIS tiles were mosaicked and re-projected from sinusoidal projection to geographic coordinates (WGS 1984) with the freely available *MRT software* (MODIS Reprojection Tool, Version 4). In particular, the *Enhanced Vegetation*

Index (EVI, MOD13A2), *Surface Reflectance* (blue, red, NIR, MIR; MOD13A2) and *Land Surface Temperature* (LST, MOD11A2) products were utilized. Pixel-level *Quality Assurance Science Data Sets* (QA-SDS) were analyzed using the TiSeG software package (Colditz et al., 2008) to exclude low-quality data, e.g. due to cloud cover or atmospheric contamination, from the time series. With a critical weighting between data quality and the necessary quantity for meaningful interpolation (Colditz et al., 2008), high-quality data were used as vertices for pixel-level linear temporal interpolation. Further, an adaptive Savitzky-Golay filter as implemented in the TIMESAT 3.0 software (Jönsson and Eklundh, 2004) was applied. The Savitzky-Golay filter is able to account for negatively-biased noise and recommended for time series with minor noise level.

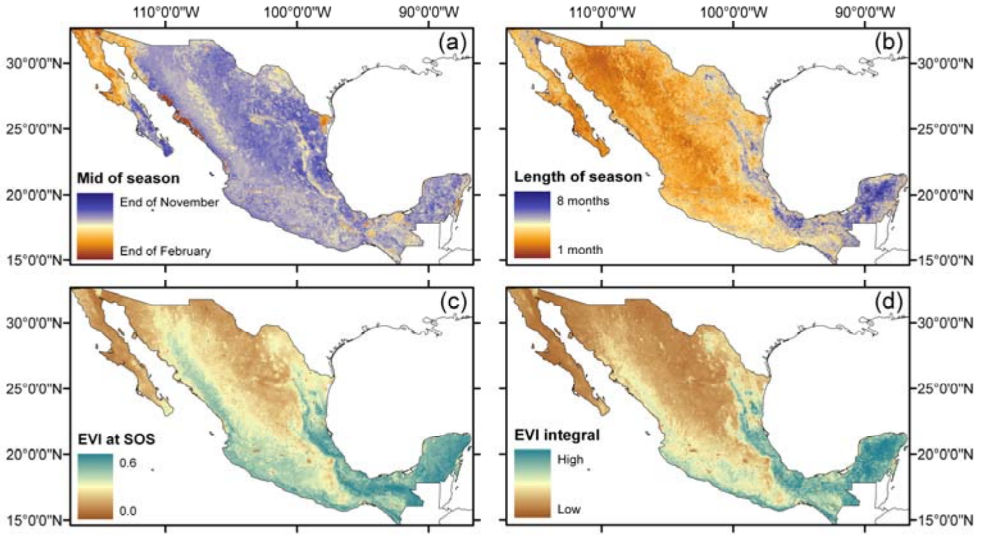

Figure 2. Selected phenological metrics derived from MODIS-EVI (*Enhanced Vegetation Index*, MOD13A2) time series. (a) Mid of season, (b) Length of season, (c) EVI value observed at the start of season, and (d) Integral under the EVI curve.

In total, 18 annual phenological metrics (Figure 2) were computed: (1) **Temporal metrics**: Start of season, mid of season, end of season, dormancy, length of season, (2) **Net primary productivity (NPP)-related metrics**: Vegetation index value at start of season, value at end of season, maximum value, minimum value, annual range, accumulated integral during vegetation period, annual mean, annual median, and (3) **Seasonality-related metrics**: rate of green-up, rate of senescence, shape of phenology curve, standard deviation, coefficient of variation. For temporal metrics referring to certain stages within the phenological cycle, the number of the corresponding composite (between 1 and 23 in accordance with the 16-day composite period of the MODIS products) was assigned. In addition, seven annual statistical metrics (minimum, mean, median, maximum, range, standard deviation, and coefficient of variation) were computed for the LST and surface reflectance time series. For each metric, the annual values were averaged over the nine years of the study period to reduce the effect of inter-annual variability.

3.4 Species distribution models

To predict suitability maps for each species, *Maximum Entropy* (Maxent) models as implemented in its software version 3.3.3e (Phillips et al., 2006) using only the previously identified non-correlated predictors were run. The models included five replicates with replicate samples selected based on bootstrap resampling (Specific settings: auto features, randomtestpoints=25, jackknife, regularization multiplier=1, maximum iterations=500, convergence threshold=0.0001).

4 RESULTS

4.1 Distribution of land cover classes observed at species presence sites

The distribution of land cover classes observed at the presence localities was assessed (Figure 3). Accordingly, the study species showed different frequency distributions of the land cover classes found at the respective presence sites. As illustrated in Figure 3a, 85.4% (374) of the presence sites of *Astronium graveolens*, typically occurring in tropical or sub-tropical broadleaf evergreen forest (see Table 1), were found in the same corresponding land cover class. All other land cover classes were represented with significantly lower frequencies. For *Liquidambar macrophylla*, only 45.0% of the records were occurring in the same most important land cover class (*Mixed forest*). For this species, the remaining presence sites were distributed over several other classes with comparatively high frequencies (Figure 3b). The distribution of land cover classes observed at the presence sites differed largely between all study species.

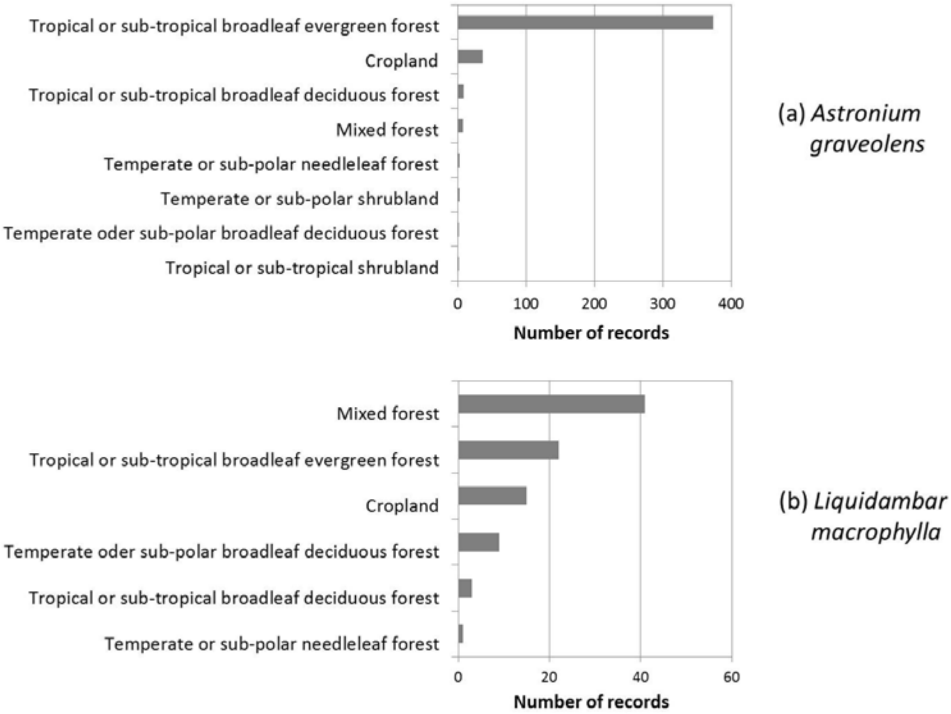

Figure 3. Exemplary frequency distribution of land cover classes observed at the presence localities of the study species. (a) *Astronium graveolens* and (b) *Liquidambar macrophylla*.

4.3 Model performance

Model accuracy was assessed based on the *area under curve* (AUC) which is calculated by summing the area under the *receiver operating characteristic* (ROC) plot. As summarized in Table 2, all models produced – according to the classification of Swets (1988) – ‘fair’ to ‘excellent’ model accuracies measured by AUC. Both training and test AUC were higher for the remote sensing data based model except for *Pinus chiapensis* with a slightly higher test AUC score for land cover data (though with a very high standard deviation). However, there was considerable variation in AUC scores between species.

A very similar trend was found for the statistical model deviance from reference presence-absence records (Table 3; calculated as implemented in the R package ‘dismo’) with higher deviance for all models developed from land cover data except for the species *Avicennia germinans*. The reduction in model deviance of the remote sensing-based as opposed to the land cover-based models was highest for

Abies religiosa (-86.3%), *Liquidambar styraciflua* (-86.2%), and *Liquidambar macrophylla* (-80.7%). The lowest scores were ascertained for *Avicennia germinans* (increase in model deviance of +10.5%), *Arbutus xalapensis* (-6.7%), and *Bursera simaruba* (-10.6%).

Table 2. Comparison auf AUC and standard deviation (SD) scores of the land cover and remote sensing based models. Higher values are printed in bold.

	Land Cover				***Remote Sensing***			
	Training		***Test***		***Training***		***Test***	
Species	***AUC***	***SD***	***AUC***	***SD***	***AUC***	***SD***	***AUC***	***SD***
Abies religiosa	0.888	0.011	0.886	0.013	**0.991**	0.002	**0.981**	0.003
Alnus acuminata	0.819	0.010	0.804	0.020	**0.955**	0.003	**0.928**	0.019
Arbutus xalapensis	0.832	0.005	0.834	0.004	**0.875**	0.002	**0.855**	0.003
Astronium graveolens	0.871	0.009	0.858	0.013	**0.938**	0.003	**0.919**	0.009
Avicennia germinans	0.979	0.005	0.963	0.037	**0.988**	0.005	**0.976**	0.015
Bursera bipinnata	0.852	0.014	0.839	0.024	**0.964**	0.002	**0.942**	0.009
Bursera simaruba	0.824	0.002	0.821	0.004	**0.866**	0.002	**0.857**	0.004
Cedrela odorata	0.821	0.007	0.809	0.014	**0.917**	0.006	**0.871**	0.002
Guaiacum sanctum	0.864	0.010	0.876	0.028	**0.971**	0.004	**0.938**	0.015
Liquidambar macrophylla	0.739	0.013	0.749	0.054	**0.957**	0.007	**0.935**	0.031
Liquidambar styraciflua	0.745	0.020	0.741	0.035	**0.977**	0.003	**0.949**	0.020
Pinus chiapensis	0.820	0.039	**0.802**	0.140	**0.932**	0.010	0.801	0.059

Table 3. Comparison auf statistical model deviance scores of the land cover and remote sensing based Maxent models. Lower values are printed in bold. % difference refers to the reduced (-) or increased (+) model deviance of the remote sensing-based as opposed to the land cover-based models.

Species	***Model deviance (land cover)***	***Model deviance (remote sensing)***	***% difference***
Abies religiosa	0.409	**0.056**	-86.3
Alnus acuminata	0.641	**0.252**	-60.7
Arbutus xalapensis	0.659	**0.615**	-6.7
Astronium graveolens	0.430	**0.314**	-27.0
Avicennia germinans	**0.038**	0.042	+10.5
Bursera bipinnata	0.520	**0.249**	-52.1
Bursera simaruba	0.739	**0.661**	-10.6
Cedrela odorata	0.647	**0.429**	-33.7
Guaiacum sanctum	0.446	**0.171**	-61.7
Liquidambar macrophylla	0.934	**0.180**	-80.7
Liquidambar styraciflua	0.899	**0.124**	-86.2
Pinus chiapensis	0.711	**0.335**	-52.9

5 DISCUSSION

5.1 Overestimation of species distributions with land cover data

Models based on land cover data tended to overestimate species distribution ranges as no continuous geographic variation or floristic gradients were represented or evident from these categorical data. Land cover data further typically suffer from cartographic generalization and often lack sufficient spatial (Kerr and Ostrovsky, 2003) and thematic (Jönsson and Eklundh, 2004) detail. For example, the difference in model training AUC between remote sensing and land cover based models was highest for the two species of the cloud forest (*Liquidambar* spp.). Cloud forest is not included as separate land cover class in the legend of the NALCMS data set (Figure 1). In this case, the failure of land cover data to model the species distributions is hence an indicator of insufficient thematic detail.

5.2 Species characteristics

The general trend towards overestimation of species distribution ranges based on categorical land cover information was found to be influenced as well by the specific characteristics of the target species. In general, consistent with the findings of Hernandez et al. (2006), higher AUC scores for both remote sensing-based and land cover-based models were found for species with small sample sizes and hence limited geographical ranges (Tables 1 and 3). In addition, the performance of land cover data for modelling species distributions was dependent on how closely the spatial distribution patterns of a species could be linked to certain land cover types. The geographical distribution of the mangrove species *Avicennia germinans* (the only species with lower model deviance scores for the land cover-based than the remote-sensing-based model) was characterized by only one dominant land cover class, namely *Wetland*. The proportion of this land cover class in relation to the Mexican land surface was only 1.0%; the class *Wetlands* could further be mapped with very high accuracies (User's Accuracy 96.4%; Colditz et al., 2010). Consistent with this, *A. germinans* showed a very low improvement in AUC (training: 0.009; test: 0.013; Table 2) scores compared to the other study species. *A. germinans* was hence the only species for which the use of remote sensing data did not improve the Maxent species distribution model. On the contrary, the highest increase in AUC scores and at the same time decrease in model deviance was observed for *Liquidambar macrophylla* (-0.232; -80.7%) and *Liquidambar styraciflua* (-0.218; -86.2%). For both species, the land cover class *Mixed forest* was the most important category observed at their presence sites. The increase in AUC and reduction in model deviance due to the use of remote sensing instead of land cover data was presumably the result of the comparatively low mapping accuracy of the class *Mixed Forest* (Producer's Accuracy: 80.1%, User's Accuracy: 62.9%; Colditz et al., 2010) and the insufficient significance of the class definitions for the target species. Both *L. styraciflua* and *L. macrophylla* occur in tropical montane cloud forests which are not represented in the legend of the NALCMS land cover product but can be characterized based on continuous multi-temporal remote sensing data.

6 CONCLUSION

To summarize, the suitability of each land cover product to predict species distributions is based on the detail (number of classes) and validity (significance of the class definitions to characterize the biotope requirements of the target species) of its legend. Further, the qualification of a certain land cover product is dependent on (1) the distribution of the land cover classes observed at the species presence sites, (2) the proportions of the study area that are covered by the respective most important land cover class(es), and (3) the mapping accuracy of the dominant land cover class(es) observed at the majority of species presence sites. In view of the generally higher mapping accuracies and greater thematic detail of regional land cover data, the use of regional or even continental rather than global land cover products is therefore recommended in species distribution modelling. Even though this analysis was conducted only for a specific land cover product, similar results can be expected for other land cover classifications. Since the pre-processing effort of remote sensing data is high compared to often readily available categorical land cover data, a trade-off situation is created between target model accuracy and processing effort.

REFERENCES

Anderson, R. P., D. Lew, and A. T. Peterson, Evaluating predictive models of species' distributions: Criteria for selecting optimal models. *Ecological Modelling 162*, 211–232, 2003.

Bradley, B. A. and E. Fleishman, Can remote sensing of land cover improve species distribution modelling? *Journal of Biogeography 35*, 1158–1159, 2008.

Colditz, R. R., C. Conrad, T. Wehrmann, M. Schmidt, and S. Dech, TiSeG: A flexible software tool for time-series generation of MODIS data utilizing the Quality Assessment Science Data Set. *IEEE Transactions on Geoscience and Remote Sensing 46*(10), 20–30, 2008.

Colditz, R. R., P. Maeda, G. López, I. Cruz, and R. Ressl, Land cover classification of Mexico in the framework of the North American Land Change Monitoring System. In *Proceedings of 2010 ASPRS Annual Conference, April 26-30, 2010, San Diego, USA*.

CONAFOR, Inventario Nacional Forestal y de Suelos, Manual y procedimientos para el muestreo de campo (RE-MUESTREO 2009). Technical report, Comision Nacional Forestal, Mexico-City, Mexico, 2009.

Heikkinen, R. K., M. Luoto, M. B. Araújo, R. Virkkala, W. Thuiller, and M. T. Sykes, Methods and uncertainties in bioclimatic envelope modelling under climate change. *Progress in Physical Geography 30*(6), 751–777, 2006.

Hernandez, P. A., C. H. Graham, L. L. Master, and D. L. Albert, The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography 29*, 773–785, 2006.

Jönsson, P. and L. Eklundh, TIMESAT – A program for analyzing time-series of satellite sensor data. *Computers & Geosciences 30*, 833–845, 2004.

Kerr, J. and M. Ostrovsky, From space to species: Ecological applications for remote sensing. *TRENDS in Ecology and Evolution 18(6)*, 299–305, 2003.

Lambin, E. F., Monitoring forest degradation in tropical regions by remote sensing: Some methodological issues. *Global Ecology and Biogeography 8*, 191–198, 1999.

Luoto, M., R. Virkkala, and R. K. Heikkinen, The role of land cover in bioclimatic models depends on spatial resolution. *Global Ecology and Biogeography 16*, 34–42, 2007.

Mateo, R. G., T. B. Croat, A. M. Felicísimo, and J. Muñoz, Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections. *Diversity and Distributions 16*, 84–94, 2010.

NALCMS (2005). North American Land Cover at 250 m spatial resolution. Produced by: Natural Resources Canada / Canadian Centre for Remote Sensing (NRCan/CCRS), United States Geological Survey (USGS), Insituto Nacional de Estadística y Geografía (INEGI), Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO), and Comisión Nacional Forestal (CONAFOR), available from: www.cec.org/naatlas/nalcms/.

Phillips, S. J., R. P. Anderson, and R. E. Schapire, Maximum entropy modeling of species geographic distributions. *Ecological Modelling 190*, 231–259, 2006.

Ricker, M., I. Ramírez-Krauss, G. Ibarra-Manríquez, E. Martínez, C. H. Ramos, G. González-Medellín, G. Gómez-Rodríguez, J. L. Palacio-Prieto, and H. M. Hernández, Optimizing conservation of forest diversity: A country-wide approach in Mexico. *Biodiversity and Conservation 16*, 1927–1957, 2007.

Sarukhán, J., P. Koleff, J. Carabias, J. Soberón, R. Dirzo, J. Llorente-Bousquets, G. Halffter, R. González, I. March, A. Mohar, S. Anta, and J. de la Maza, *Natural capital of Mexico. Synopsis: Current knowledge, evaluation, and prospects for sustainability*. Comisión Nacional para el Conocimiento y Uso de la Biodiversidad, Mexico-City, Mexico, 2010.

Swets, J., Measuring the accuracy of diagnostic systems. *Science 240*, 1285–1293, 1988.