Environ Monit Assess (2012) 184:5439–5451
DOI 10.1007/s10661-011-2351-9

Regional data refine local predictions: modeling the distribution of plant species abundance on a portion of the central plains

Nicholas E. Young · Thomas J. Stohlgren · Paul H. Evangelista · Sunil Kumar · Jim Graham · Greg Newman

Received: 26 April 2011 / Accepted: 30 August 2011 / Published online: 13 September 2011

Abstract Species distribution models are frequently used to predict species occurrences in novel conditions, yet few studies have examined the consequences of extrapolating locally collected data to regional landscapes. Similarly, the process of using regional data to inform local prediction for species distribution models has not been adequately evaluated. Using boosted regression trees, we examined errors associated with extrapolating models developed with locally collected abundance data to regional-scale spatial extents and associated with using regional data for predictions at a local extent for a native and non-native plant species across the northeastern central plains of Colorado. Our objectives were to compare model results and accuracy between those developed locally and extrapolated regionally, those developed regionally and extrapolated locally, and to evaluate extending species distribution modeling from predicting the probability of presence to predicting abundance. We developed models to predict the spatial distribution of plant species abundance using topographic, remotely sensed, land cover and soil taxonomic predictor variables. We compared model predicted mean and range abundance values to observed values between local and regional. We also evaluated model prediction performance based on Pearson's correlation coefficient. We show that: (1) extrapolating local models to regional extents may restrict predictions, (2) regional data can help refine and improve local predictions, and (3) boosted regression trees can be useful to model and predict plant species abundance. Regional sampling designed in concert with large sampling frameworks such as the National Ecological Observatory Network may improve our ability to monitor changes in local species abundance.

Keywords Boosted regression trees · Species distribution models · Spatial scale · Extrapolation · National Ecological Observatory Network · Abundance

N. E. Young (✉) · P. H. Evangelista · S. Kumar · J. Graham · G. Newman
Natural Resource Ecology Laboratory,
Colorado State University,
Fort Collins, CO, USA
e-mail: neyoung@rams.colostate.edu

T. J. Stohlgren
USGS, Fort Collins Science Center,
2150 Centre Ave,
Fort Collins, CO, USA

Introduction

Scaling ecological patterns and processes is a lingering challenge for ecologists (Levin 1992). Using the appropriate scale can have impacts on the ability to not only identify ecological patterns but also to understand the processes driving those patterns (Scott et al. 2002). For example, the drivers of change at a local scale are often influenced by past disturbances

while drivers at a continental scale are primarily climatic (Brown et al. 2008). One area where scale has been especially challenging is predicting species distributions; often with the use of species distribution models (SDM).

Species distribution models relate species response data (either occurrence or abundance) with environmental characteristics (Elith and Leathwick 2009). These models have met many management objectives including identifying previously unknown populations of endangered species (Evangelista et al. 2008b), predicting vulnerable habitats to species invasions (Stohlgren et al. 2002), estimating species richness (Graham and Hijmans 2006), and many others (Elith and Leathwick 2009). The breadth and application of SDMs has rapidly increased in the past decade. Advances in computer capabilities, availability of geospatial environmental data, and powerful geographic information systems facilitate modeling complex ecological interactions to predict species distributions (Guisan and Thuiller 2005; Elith and Leathwick 2009). Some of the most common SDMs include Maxent (Phillips et al. 2006), boosted regression trees (BRT; Friedman et al. 2000), multivariate adaptive regression splines (Friedman 1991), and Random Forests (Breiman 2001). Each algorithm offers strengths and weaknesses for modeling species distributions; multiple studies compare these methods (Araujo and New 2007; Elith and Graham 2009; Kumar et al. 2009; Parisien and Moritz 2009). Although these models are often compared using the same data set, environmental predictors, and spatial scale, it is important to consider that SDMs are designed to handle different types of data sets and perform best under specific circumstances. For example, some models are designed for datasets with only presence locations like Biomapper (Hirzel et al. 2002) or Maxent while other models such as BRT or Random forests require presence and absence data.

Although SDMs were developed to model species within the environment from which the data were collected (Guisan and Zimmermann 2000), these models are now being used to predict species distributions in space and/or time to novel conditions not representative of the modeled environment (Elith et al. 2010). This has been referred to as model projection, generalization, transfer, and extrapolation (Fielding and Haworth 1995; Randin et al. 2006), hereafter referred to as extrapolation. Model extrapolation has been used to predict the distribution of a species under climate change (Penman et al. 2010), in hypothesized susceptible regions of invasion (Medley 2010), and over large extents (Mateo-Tomas and Olea 2010). While insights may be gained through extrapolation, recent studies have suggested that this may not always be the best approach to model species distributions in novel environments (Pearson et al. 2006). Thuiller et al. (2004) demonstrated that extrapolating SDMs to regions not representing the complete range of environmental conditions may lead to highly liberal predictions of species occurrence. On the other hand, regional data are more difficult to collect and require more resources. Therefore, valuable resources (e.g., personnel, time, and money) will be saved if local data can be extrapolated to make accurate regional predictions.

While there have been some studies looking at extrapolating local prediction to regional extents, to our knowledge there have not been any studies that have looked at the impact of using regional data to make predictions at a local extent. In most cases, the extent of interest (e.g., national park, watershed, political boundary, etc.) is modeled using data collected at that extent (e.g., Stohlgren et al. 2010). This makes sense when modeling the global distribution of a species, but becomes questionable when modeling a portion of a species distribution. Here, we explore this concept and hypothesize that using regional data to make predictions at a local extent can help capture larger environmental and response variation which has been shown to improve model results (Elith and Graham 2009) and improve local predictions.

Most SDMs are designed exclusively for presence-only or presence–absence data and are not compatible with abundance data. Under this framework, these models predict the probability of presence or probability of suitable habitat rather than a prediction of the number of species. Estimates of abundance allow managers to prioritize management actions that may not be possible with predicted presence alone. Before the advent of powerful computers and geographic information systems, previous studies predicting species abundance primarily used regression methods (Evangelista et al. 2004; Crall et al. 2006). These methods still serve as a foundation to many recently developed models (e.g., random forests and boosted regression trees). Identifying and predicting the spatial pattern of species abundance has advanced

through the increased use of geographic information systems and spatial models (Sagarin et al. 2006). Part of the issue surrounding the lack of distribution models predicting abundance in the ecological community is the fact that modeling abundance data requires more robust statistical models than presence–absence data (Austin 2002) and abundance data are inherently rarer than presence/absence or presence-only data. Still, there are some recent examples where SDM were used to model predicted abundance on a landscape (e.g., Strubbe et al. 2010). The accuracy and predictive capabilities of models that extrapolate the distribution of species abundance from local to regional extents using only locally collected data are largely unknown.

From management perspective, non-native species continue to be an economic burden for organizations responsible for maintaining ecosystem integrity and processes (Mack et al. 2000). In most cases, the task of surveying an entire area for non-native species is unrealistic. Providing spatial abundances models of non-native species can help managers spatially prioritize detection, control and prevention efforts. The purpose of this study was to examine the issue of scale (in terms of spatial extent) and abundance predictions associated with extrapolating SDMs developed using locally collected data to regional extents and predictions associated with using regional data to make predictions at the local extent. Specifically, our objectives were to: (1) investigate how extrapolating from local to regional scales impacts model results and accuracy (2) compare local extent model results and accuracy from models developed using data inside the local extent and to those developed with data inside and outside the local extent, (3) evaluate extending SDM from predicting probability of presences to predicting abundance using boosted regression trees. We used abundance data (i.e., percent foliar cover) of a native and non-native species on a portion of the central plains and boosted regression trees.

Methods

Study area

We examined data at two extents on the central plains of eastern Colorado (Fig. 1). We chose these extents because they represent two of the four strategic designs the National Ecological Observatory Network (NEON) has identified in their continental-scale research platform for discovering and understanding the impacts of climate change, land-use change, and non-native species on ecosystem processes (NEON 2010). There are 20 ecoclimatic domains established by NEON across the USA, and the local and regional extents in this study represent the core wildland site and the approximate footprint of aerial observations (airborne observatory platform), respectively, within the Central Plains domain (domain 10).

The core wildland site for domain 10 is at the Central Plains Experimental Range, which is located in the Colorado Piedmont section of the Great Plains (40°49′ N and 104°46′ W). Covering 6,798 ha, the site represents the local extent of our study area and focal point of our research. This area is a semi-arid, C_4-dominated native shortgrass steppe ecosystem. Most of the precipitation occurs during the growing season from April to September. Grazing by domestic cattle is the dominant land use in conjunction with research and monitoring projects including prescribed fire (Shortgrass Steppe Long Term Ecological Research; http://www.sgslter.colostate.edu).

The regional extent of our study area covers an area of 40,000 ha, which represents the 20×20 km footprint of aerial observations defined by NEON for this domain. This area was defined by NEON to collect detailed aerial data about land-use and vegetation structure (NEON). The regional extent encompasses the local extent and is similar in terms of climate and ecological characteristics; however, land use is more diverse. Much of the regional extent is a mosaic of shortgrass steppe, agricultural land, rangeland, and human development. Highway 85 runs north–south down the middle of the regional extent and the large developed area surrounding the town of Nunn is located in the southern portion.

Species

We chose to model common native and non-native species to the central plains. *Bouteloua gracilis* (blue grama), a warm season perennial bunchgrass, is native to the shortgrass steppe and considered a dominant species throughout the Colorado Piedmont. Although it has evolved on the shortgrass steppe, Marilyn and Hart (1994) found that *B. gracilis* recovered poorly on disturbed sites. *Sisymbrium altissimum* (tall tumble-

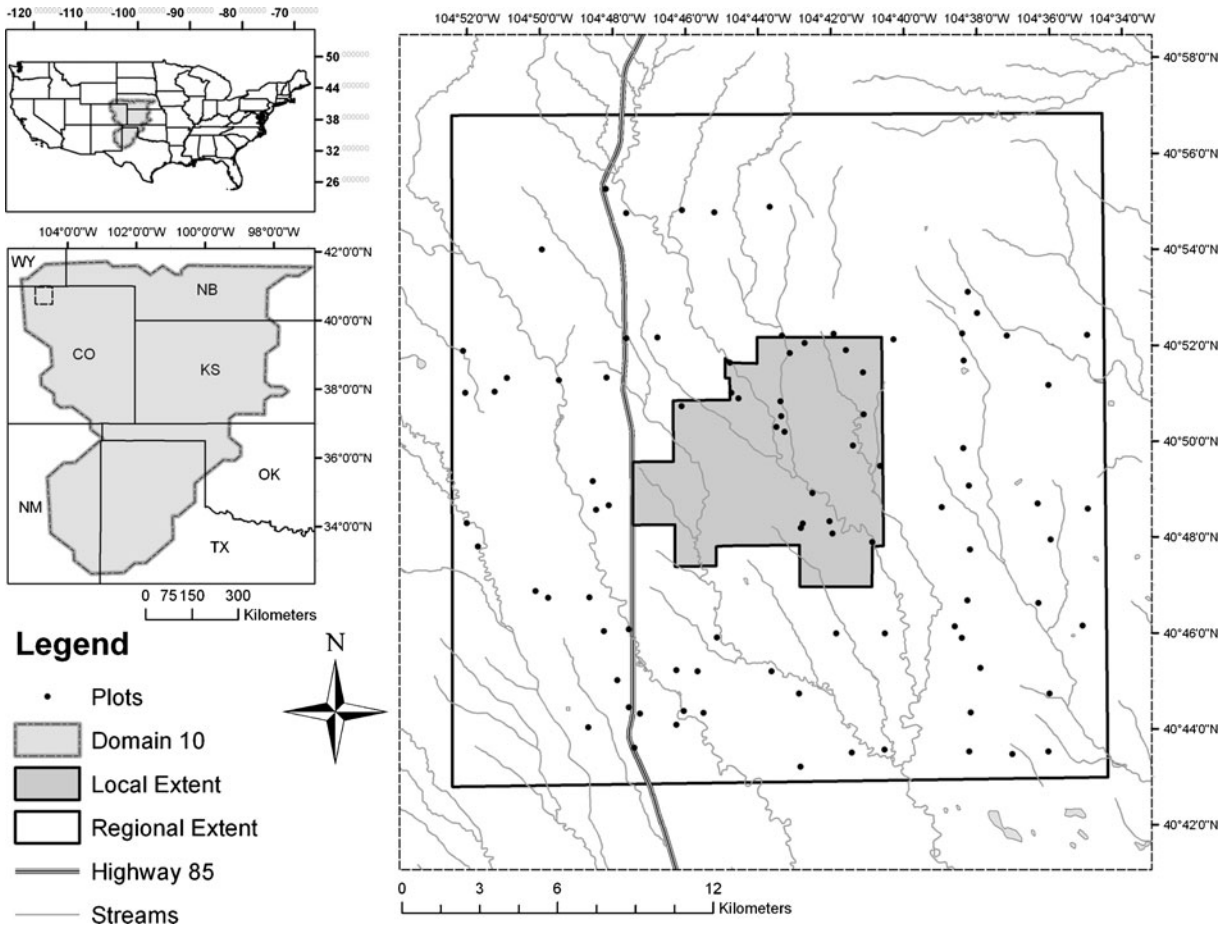

Fig. 1 Study area showing sampled plots at the local and regional extents. The local and regional extents are within the larger central plains domain (National Ecological Observatory Network domain 10)

mustard) is a non-native annual or biannual species found in disturbed sites with other non-native and native annuals (Allen and Knight 1984). *S. altissimum* can be found on many different soil types (Patman and Hugh 1961). Although both species are commonly found in the shortgrass steppe ecosystem, *B. gracilis* is generally a dominant species, while *S. altissimum* is rarely dominant.

Field data

Local and regional abundance data were combined from two separate studies. Data for the local extent were collected in 2008 as a part of a NEON preliminary assessment for the Central Plains Experimental Range (Evangelista et al. 2009a) that consisted of 20 sampled plots. Vegetation cover abundance data were recorded by estimating the percent cover within a 168-m^2 circular, multi-scale vegetation plot modified from the National Forest Service Inventory and Analysis Program (Barnett et al. 2007); (Frayer and Furnival 1999). Regional abundance data were collected with 72 Braun-Blanquet plots (Braun-Blanquet 1932). This relatively quick method of sampling is suited for species–environment relationships (Wikum and Shanholtzer 1978). These data were collected at the regional extent with the exception of two locations which were sampled inside the local extent. These two samples were added to the local extent dataset. This provided a dataset of 92 samples for the regional models and 22 samples for the local models.

Environmental variables

We included soil, land cover, topographic and remotely sensed environmental data as our predictor varia-

bles (Table 1). All predictor variables had a 30-m resolution. Topographic variables consisted of elevation, slope, aspect, and solar radiation. No climatic variables were included in the models because the spatial extent was too small for these predictors to be important drivers.

Soil data were downloaded from Soil Data Mart provided by US Department of Agriculture Natural Resource Conservation Service (http://soildatamart.nrcs.usda.gov/). Soil texture has been shown to be an important predictor of individual plant species on the shortgrass steppe (Hook et al. 1991). These data were originally classified by map unit series. We classified the map series to soil great groups. Soil great groups are a classification of soil taxonomy that reflects assemblages of the horizons and the most significant properties of the whole soil (Soil Survey Staff 1999).

We downloaded LANDFIRE existing Vegetation Type land cover data from the LANDFIRE website (http://www.landfire.gov). The LANDFIRE dataset was developed using a complied field database for reference plots along with biophysical gradients and Landsat imagery (Rollins 2009). LANDFIRE uses land cover classifications defined by NatureServes's ecological systems classifications which are ecological units at mid-scale resolution. The LANDFIRE values were grouped to represent nine land cover types. Open water (11), developed (21, 22, 23, and 24), barren (31 and 2,007), agriculture (81 and 82), shrubland (2,072, 2,081, 2,086, and 2,107), grassland/forbland (2,094, 2,127, 2,181, 2,182, and 2,183), mixedgrass prairie (2,132), shortgrass prairie (2,149), and riparian (2,159 and 2,162). We used these grouped land cover types to represent classifications appropriate for the scales we were modeling and to allow for more intuitive interpretation of model results.

Six topographic predictor variables were included in the models. Topography in conjunction with soil properties can be a driving factor in plant distributions on the shortgrass steppe (Burke et al. 1999). Using a US Geological Survey 30-m digital elevation model (DEM), we calculated solar radiation in ArcGIS 9.3 (The Environmental System Research Institute, USA). We applied the time period for solar radiation calculations from 15 June 2010 to 29 June 2010 which was when the regional extent sampling occurred. Slope and aspect were also derived from the DEM and calculated using ArcGIS 9.3.

In addition to land cover and topographic variables, remotely sensed Landsat 7 ETM+ satellite scene data were downloaded for 7 July 2000 from USGS Earth Resources Observation Center (EROS, http://glovis.usgs.gov/). The scenes were the most recent cloud free images obtained when the operational scene line corrector was functioning for the season the field data were collected. The scenes and derived vegetation indices were processed using ERDAS Imagine 2010 (ERDAS Atlanta, GA, USA) and ArcGIS 9.3 software. We generated three vegetation indices: Normalized Difference Vegetation Index, Ratio Vegetation Index and Soil-Adjusted Vegetation

Table 1 List of all environmental predictors included in initial models with their data type and source

Environmental predictor	Data type	Source
Aspect	Continuous	Calculated from elevation
Elevation	Continuous	U.S. Geological Survey (http://eros.usgs.gov)
Enhanced Vegetation Index	Continuous	Calculated from Landsat bands
Greeness	Continuous	Calculated from Landsat bands
LANDFIRE vegetation class	Discrete	http://www.landfire.gov/products_national.php
Normalized Difference Vegetation Index	Continuous	Calculated from Landsat bands
Ratio Vegetation Index	Continuous	Calculated from Landsat bands
Slope	Continuous	Calculated from elevation
Soil brightness	Continuous	Calculated from Landsat bands
Soil Great Group	Discrete	Soil Data Mart (http://SoilDataMart@nrcs.usda.gov)
Soil-Adjusted Vegetation Index	Continuous	Calculated from Landsat bands
Solar radiation	Continuous	Calculated from elevation
Wetness	Continuous	Calculated from Landsat bands

Index (Li and Weng 2005). These indices are for vegetation and land cover feature estimations. Tasseled cap transformations were also conducted for the Landsat 7 scenes. Tasseled cap transformations provide measurements of soil brightness (tasseled cap, band 1), vegetation greenness (tasseled cap, band 2) and soil/vegetation wetness (tasseled cap, band 3) (Huang et al. 2002). Tasseled cap bands and vegetation indices have been shown to be effective predictors of plant occurrences when combined with SDMs (Evangelista et al. 2009b).

Analyses

For spatial analyses, we chose BRTs to model *B. gracilis* and *S. altissimum* at local and regional extents. Modeling species abundances using BRTs is a relatively new method in ecology (Elith et al. 2008). In addition to being able to model abundance data, we chose BRTs because they have been shown to perform well with small sample sizes compared with other SDMs (Wisz et al. 2008) and because they can incorporate categorical predictor variables. Boosted regression trees attempt to minimize the loss function by generalizing many simple classification and regression trees. Tree-based models, such as BRTs, accomplish this by applying rules to the predictors that partition the data into rectangles with the most homogeneous response (Elith et al. 2008). Boosting is a form of re-sampling that, unlike other methods such as bagging or sub-sampling, applies a weighted probability of a response to be re-sampled based on previous classifications (Franklin 2009). The relative importance of the predictor variables is calculated based on the number of times a predictor variable was chosen as a splitting node and weighted based on the improvement to the model based on each split (Friedman and Meulman 2003). BRTs are able to decrease over-fitting data by averaging the predictions of many trees created using subsets of the data (Franklin 2009).

We used the generalized boosted models package in R (R Development Core Team 2010) for our BRT analyses (Friedman et al. 2000). There are a few settings that can be adjusted when running BRTs. A low-learning rate decreases the model over learning but requires more iterations (De'ath 2007). Optimizing both the learning rate in conjunction with the number of trees is similar to model regularization. Regularization prevents models from over-fitting training data. Interaction depth or tree complexity is the number of nodes in each tree created. By adding more nodes to the tree, more variable interactions are added. With smaller datasets, larger tree complexity provides no advantage (De'ath 2007). We performed 5,000 iterations and ensured that at least 1,000 trees were generated for each model (Elith et al. 2008) accomplished by a learning rate of 0.001 and a tree complexity of 2.

Preliminary models for each species for both the local and regional extent were constructed using all 13 predictor variables to identify those with the greatest predictive contributions and reduce the overall number of variables in our analyses. From these results, we kept only the top 3 predictors and removed the remaining variables due to the limited sample size for local models (Maxwell 2000). From those, we performed a Pearson's cross-correlation test using SYSTAT (version 12; SYSTAT Software, Port Richmond, California, USA) to remove highly correlated variables (Pearson's correlation coefficient, >0.8 or ≤ 0.8) keeping the variable that had the highest contribution and then selected the next predictor variable with the highest relative influence to include in the final model. The variables remaining were used to develop final models for each extent. Regional models for both species were developed with data at both the local and regional extent ($n=92$) while local models were developed using only local data ($n=22$). To evaluate model performance, we used cross-validation during model development and fit a linear regression between predicted values and observed values to calculate adjusted r^2 and also calculated Pearson's correlation coefficient (Potts and Elith 2006).

Finally, we compared the mean, minimum and maximum across the extents for the predicted models. We also evaluated the difference between predicted abundance and observed abundance to compare models developed with local data and models developed using regional data. The summary statistics were calculated using SYSTAT.

Results

Predictor variables

Both the local model and the regional model for *B. gracilis* had similar predictor variables (Table 2). *S.*

Table 2 Relative influence of environmental predictors for *B. gracilis* for the local and regional models

Local		Regional	
Predictor	Relative influence	Predictor	Relative influence
Soil great group	41	Slope	43
Solar radiation	37	Soil great group	30
Slope	22	Solar radiation	27

altissimum final models for local and regional extents also had similar predictor variables (Table 3). Soil and slope proved to be key predictors for both species at both extents. Soil great group was a key contributor at the local extent models (>40% relative influence) for both species, while slope was the important contributor for the regional extent models (>40% relative influence).

Model performance

Local and regional models showed significant predictive ability for both *B. gracilis* and *S. altissimum* ($p<0.001$). The local model for *B. gracilis* had a slightly higher explained variance (adjusted $R^2=0.65$) than the regional model (adjusted $R^2=0.52$). The same was true for *S. altissimum* local (adjusted $R^2=0.45$) and regional (adjusted $R^2=0.44$) models. Local models also had a stronger association than regional models when model predictions were compared with observed values (Pearson's r for local and regional models for *B. gracilis*=0.82, 0.72, respectively, and *S. altissimum*= 0.69, 0.67, respectively).

Predictive map descriptions

For *B. gracilis* predictions at the local extent, the local model predicted more uniform abundance with higher abundance in the east closely aligned with soil taxonomy (Fig. 2a). The regional model for the local extent showed much more variation in abundance predictions than the local model, but showed similar areas of high abundance (Fig. 2b). At the regional extent, predictions for the local model of *B. gracilis* were highest in the southern portion and extended northward in bands following suitable soil types (Fig. 2c). The lowest abundance predictions were found in the northern portion of the regional extent where the land cover is more barren and includes a portion of the Pawnee Buttes National Grassland. The regional model predictions at the regional extent differs from the local model in that the highest predicted abundance values are found in a swath running from the east to the northwest (Fig. 2d) with relatively lower abundance predictions in the south and southwestern portions where there is more human development.

At the local extent, the mean abundance predictions for the local model (mean=27.4% cover, SE± 6.9) and regional model (mean=22.1% cover, SE± 7.5) for *B. gracilis* were similar, but the range of abundance for the local model (33.8% cover) was narrower than the regional model (39.0% cover; Table 4). The regional model predicted a maximum abundance of 44.4% cover which was similar to the local model maximum abundance predictions (44.8%; Table 4). Furthermore, the minimum predicted abundance for the regional model was 5.4% while the local model minimum prediction was 11.0% (Table 4).

For *S. altissimum* predictions at the local extent, the local model predicted relatively low abundance and little variation (Fig. 3a). In contrast, the regional model had greater prediction variation and predicted relatively higher abundance at the local extent

Table 3 Relative influence of environmental predictors for *S. altissimum* for the local and regional models

Local		Regional	
Predictor	Relative influence	Predictor	Relative influence
Soil great group	57	Slope	66
Soil brightness	24	Soil brightness	26
Slope	19	Soil great group	9

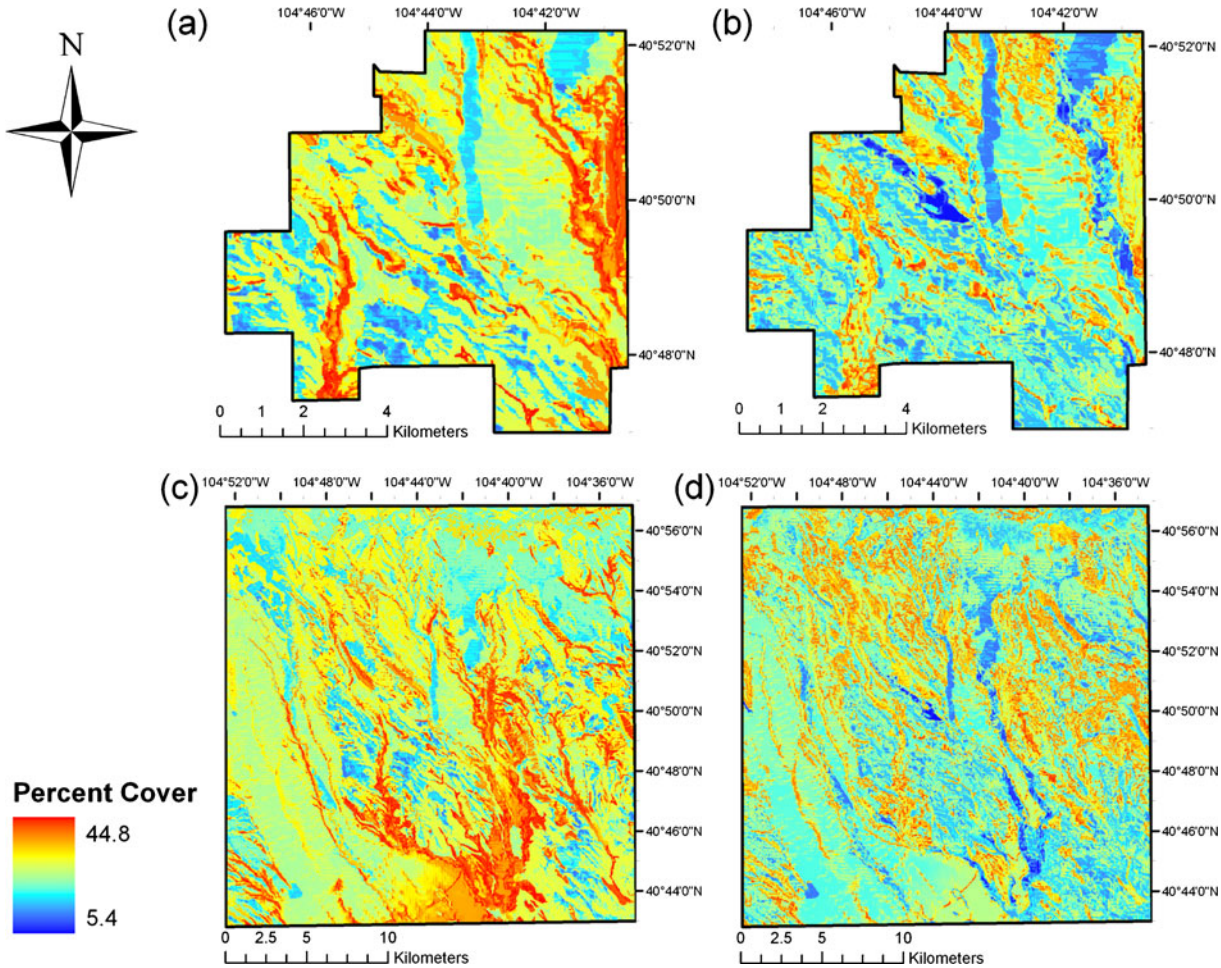

Fig. 2 Abundance predictions of *B. gracilis* local and regional models shown at two extents. **a** Local extent predictions modeled using local data and **b** local extent predictions modeled using local and regional data. **c** Regional extent predictions from the model developed using local data extrapolated to regional extent and **d** regional predictions from the model developed using local data and regional data

(Fig. 3b). At the regional extent, the local model again predicted relatively low abundance and little variation with lower abundance predictions in the southwest and southeast corners of the regional extent and higher abundance predictions in swaths that follow soil great groups (Fig. 3c). Conversely, the regional model had more variation in predictions and predicted high abundance in the south-central area of the regional extent that extended northwest (Fig. 3d). These predictions coincide with the highest observed abundance where the land use is primarily agricultural.

At the local extent, *S. altissimum* regional and local models had a large difference in predicted abundance ranges (local=0.8% cover, regional=4.1% cover). The difference in the range of predicted abundance can be attributed to the predicted maximum abundance for each model (local=1.0% cover, regional=4.7% cover) even though the predicted minimum for the local model (local=0.2% cover) was less than the regional model (regional=0.6% cover). The predicted mean of the local model was 0.5% cover (SE±0.2) and the regional model mean was 1.1% cover (SE±0.5).

Difference comparison between observe and predicted values

For *B. gracilis*, models developed using local data predicted an average 12% cover (SE±2.1, n=92)

Table 4 Local and regional model predicted maxima, minima, means, mean standard errors, and ranges across the local and regional extents

	Bouteloua gracilis		*Sisymbrium altissimum*	
	Local model predicted % cover	Regional model predicted % cover	Local model predicted % cover	Regional model predicted % cover
Regional extent				
Maximum	44.8	44.4	1.0	4.7
Minimum	11.0	5.4	0.2	0.6
Mean	28.5	25.2	0.6	1.1
Mean standard error	7.2	7.5	0.2	0.6
Range	33.8	39.0	0.7	4.0
Local extent				
Maximum	44.8	44.4	1.0	4.7
Minimum	11.0	5.5	0.2	0.7
Mean	27.4	22.1	0.5	1.1
Mean standard error	6.9	7.5	0.2	0.5
Range	33.8	39.0	0.8	4.1

lower than observed values and 0.4% lower (SE±0.4, n=92) for *S. altissimum*. Conversely, the regional models were off by an average 7% cover (SE±2.1, n=92) for *B. gracilis* and 0.1% cover (SE±0.3, n=92) for *S. altissimum*. For both species, the regional models predicted abundance values closer to the observed values than the local models.

Discussion

Using regional data can refine local predictions

Incorporating regional data improved the model range of predictions at local and regional extents. For both species, the models developed using regional data to model the local geographical extent showed a larger range of predicted abundances (Figs. 2d and 3d). This was especially true for *S. altissimum* where the maximum predicted abundance for the regional model was more than four times that of the local model. The improved predictions may be attributed to the additional landscape elements included by increasing the extent sampled (Wiens 1989).

A larger range of predicted abundance can provide more detail and easier interpretations of predictions. Our results suggest the importance of collecting data outside the local area to not only capture environmental variation, but also species response variation. For example, locations with higher abundance of a species of interest might exist just outside the area of interest being modeled; excluding data outside the area increases the possibility of an invasion going unnoticed. Modeling the potential distribution of invaders in the local area is essential to non-native species risk characterization (Stohlgren and Schnase 2006), and only possible by sampling outside the local area. Collecting additional regional samples may improve model predictions and reveal patterns missed by local models.

Extrapolation may restrict predictions

Using BRTs to extrapolate models using local data to regional extents may constrict the range of predicted values. Our results show that when extrapolating local models to regional extents, predictions in the regional extent will not exceed the range of predictions within the local extent. For both *S. altissimum* and *B. gracilis*, the models developed using data collected at the local extent did not predict abundance values below the minimum or above the maximum predicted within the local extent (Figs. 2 and 3). Our results are similar to those of Menke et al. (2009) who looked at extrapolation of an Argentine ant in southern California and found if predictions are to be made for larger un-sampled regions, additional sampling is needed to capture the environmental variation in those regions. Similarly, Randin et al. (2006) found

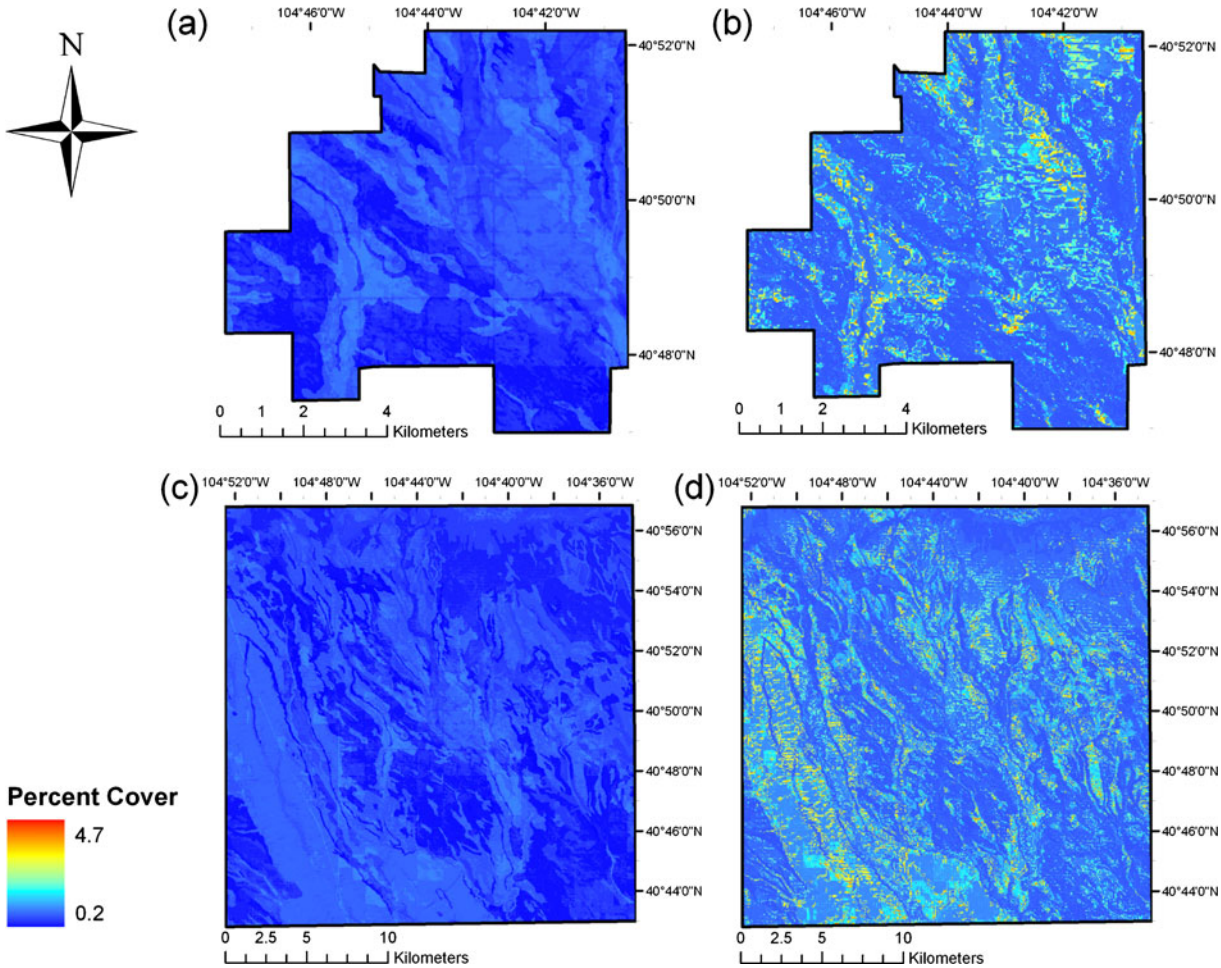

Fig. 3 Abundance predictions of *S. altissimum* local and regional models shown at two extents. **a** Local extent predictions modeled using local data and **b** local extent predictions modeled using local and regional data. **c** Regional extent predictions from the model developed using local data extrapolated to regional extent and **d** regional predictions from the model developed using local data and regional data

restricted predictions when extrapolating to another region. How a model will predict when extrapolated to novel environments appears to depend on the specific model implemented (Pearson et al. 2006). Boosted regression trees fit response curves for each predictor and, for environmental values outside the sample variation, the response curves remain constant (Elith and Graham 2009). This explains why the abundance range for models developed using only local data was the same for both the local and regional extent. Other evaluations of SDM extrapolation reveled that models may over predict or under predict when extrapolated to novel conditions (Peterson et al. 2007). Thuiller et al. (2004) found that training a model using limited environmental conditions may cause unpredictable effects on the tails of the response curves leading to poor extrapolations to wider ranges of environmental conditions.

The uncertainty surrounding model extrapolation will continue to prompt new studies that explore ways to improve model extrapolations. For example, Miller et al. (2004) recommend using simple mechanistic relationships that are well understood when extrapolating beyond narrow ranges. Elith et al. (2010) suggested smoothing the initial models to improve fitting a model to the species rather than the specific data set when the model will be extrapolated. Others have also recommended using ensemble modeling (Araujo and New 2007) or improving model calibration (Phillips and Elith 2010).

Modeling abundance using boosted regression trees

Our results show BRTs can be an effective method to model plant species abundance. We generated accurate spatial distribution models for plant species abundance at both local and regional extents. While the use of BRTs to predict abundance appears promising, uncertainty is inherent to all models and results should be carefully interpreted (Elith and Leathwick 2009). More specifically, decision trees are sensitive to the response data and environmental predictors being modeled (Berk 2008). Modifying either of these can result in very different models that have similar measured predictive abilities (Scull et al. 2005). Interestingly, the *S. altissimum* regional model preformed the poorest of all four models. This may be because *S. altissimum* is a generalist species which tend to be more difficult to predict (Evangelista et al. 2008a) or because of the relatively low abundance of *S. altissimum* (often only observed at 1% cover). While *S. altissimum* was generally observed at low abundance values, a few locations were observed to have over 25% cover. These high abundances are not common but are important to recognize for non-native species management, and our results suggest that BRTs may not predict abundances that are unusually high. Also of note, none of the models predicted abundance at 0% while there were many sampled plots that had 0% cover. This may indicate the predications should be interpreted as more of a relative abundance than absolute. Therefore, modeled abundance predictions may be best applied in conjunction with model probability of presence. The probability of presence models can be used to identify locations of interest from which the abundance predictions can provide a measure of the biomass at those locations.

The ability to predict abundance rather than just probability of presence may provide more than just where a species may occur but also information on the quality of habitat (Pearce and Ferrier 2001). In terms of non-native species, this information may help managers identify possible susceptible life stages to control and prevent invasion (Brown et al. 2008). When managing and monitoring non-native species, abundance predictions can help prioritize control and prevention efforts in addition to early detection. Many SDMs are limited to presence–absence or presence-only data. This is most likely due to the costs associated with obtaining abundance data compared with presence–absences or presence-only data. This has prompted comparison studies that investigated possible correlations between probability of presence and abundance. Unfortunately, these studies found little correlation, and if so, only between high probability of occurrence and high abundance (Pearce and Ferrier 2001; but see Vanderwal et al. 2009).

Conclusions

Extrapolating local models to regional extents is likely to predict abundance further from observed values when compared with models that included regional data. Model extrapolation in time or space is prone to violating assumptions of SDMs (Wiens et al. 2009). This is especially true for non-native species because they are rarely at equilibrium with the environment. When possible, additional samples should be collected in the regional extent to improve predictions. These additional data can provide insights into populations that may be just outside the local extent and would otherwise go unnoticed. This information can be important for regional and local conservation planning in prioritizing management efforts. Future work may investigate the number of additional regional samples required and their optimal location to provide the best predictions. Boosted regression trees can be a useful tool for modeling, and while BRTs are still largely unused in ecology (De'ath 2007), the recent increase of BRTs in the literature is promising. The ability to take advantage of BRT capabilities will depend on abundance data available with the development of large databases collecting and disseminating ecological data (Graham et al. 2007). An iterative approach to surveys and modeling may gain a more comprehensive understanding of modeling abundances and possible errors stemming from model extrapolation. More research into the use and extrapolation of species distribution models to predict species abundance is needed to fully understand the errors and benefits of this approach.

Acknowledgments The research for this study was funded by The US Geological Survey (USGS), the National Science Foundation and the National Ecological Observatory Network, Inc. We would like to thank the Natural Resource Ecology Laboratory at the Colorado State University for facility use and expertise. In addition, we would like to thank the Colorado State University herbarium and staff for access to their collection and for their expertise.

References

Allen, E. B., & Knight, D. H. (1984). The effects of introduced annuals on secondary succession in sagebrush-grassland, wyoming. *The Southwestern Naturalist, 29*, 407–421.

Araujo, M. B., & New, M. (2007). Ensemble forecasting of species distributions. *Trends in Ecology & Evolution, 22*, 42–47.

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

Barnett, D., Stohlgren, T., Jarnevich, C., Chong, G., Ericson, J., Davern, T., et al. (2007). The art and science of weed mapping. *Environmental Monitoring and Assessment, 132*, 235–252.

Berk, R. A. (2008). *Statistical learning from a regression perspective*. New York: Springer.

Braun-Blanquet, J. (1932). *Plant sociology. The study of plant communities*. New York: McGraw-Hill.

Breiman, L. (2001). Random forests. Technical report 567, University of California Berkeley, CA.

Brown, K. A., Spector, S., & Wu, W. (2008). Multi-scale analysis of species introductions: combining landscape and demographic models to improve management decisions about non-native species. *Journal of Applied Ecology, 45*, 1639–1648.

Burke, I. C., Lauenroth, W. K., Riggle, R., Brannen, P., Madigan, B., & Beard, S. (1999). Spatial variability of soil properties in the shortgrass steppe: the relative importance of topography, grazing, microsite, and plant species in controlling spatial patterns. *Ecosystems, 2*, 422–438.

Crall, A. W., Newman, G. J., Stohlgren, T. J., Jarnevich, C. S., Evangelista, P., & Guenther, D. (2006). Evaluating dominance as a component of non-native species invasions. *Diversity and Distributions, 12*, 195–204.

De'ath, G. (2007). Boosted trees for ecological modeling and prediction. *Ecology, 88*, 243–251.

Elith, J., & Graham, C. H. (2009). Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. *Ecography, 32*, 66–77.

Elith, J., & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics, 40*, 677–697.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology, 77*, 802–813.

Elith, J., Kearney, M., & Phillips, S. (2010). The art of modelling range-shifting species. *Methods in Ecology and Evolution, 1*, 330–342.

Evangelista, P., Stohlgren, T. J., Guenther, D., & Stewart, S. (2004). Vegetation response to fire and postburn seeding treatments in juniper woodlands of the grand staircase-Escalante national monument, Utah. *Western North American Naturalist, 64*, 293–305.

Evangelista, P. H., Kumar, S., Stohlgren, T. J., Jarnevich, C. S., Crall, A. W., Norman, J. B., et al. (2008). Modelling invasion for a habitat generalist and a specialist plant species. *Diversity and Distributions, 14*, 808–817.

Evangelista, P. H., Norman, J., Berhanu, L., Kumar, S., & Alley, N. (2008). Predicting habitat suitability for the endemic mountain nyala (*Tragelaphus buxtoni*) in Ethiopia. *Wildlife Research, 35*, 409–416.

Evangelista, P., Barnett, D., Stohlgren, T.J., Stapp, P., Jarnevich, C., Kumar, S. & Rauth, S. (2009a). Field and costs assessment for the fundamental sentinel unit (FSU) at the central plains experimental range, Colorado. Technical Report for National Ecological Observatory Network, Inc., Boulder

Evangelista, P., Stohlgren, T., Morisette, J., & Kumar, S. (2009). Mapping invasive tamarisk (*Tamarix*): a comparison of single-scene and time-series analyses of remotely sensed data. *Remote Sensing, 1*, 519–533.

Fielding, A. H., & Haworth, P. F. (1995). Testing the generality of bird-habitat models. *Conservation Biology, 9*, 1466–1481.

Franklin, J. (2009). *Mapping species distributions*. Cambridge: Cambridge University Press.

Frayer, W. E., & Furnival, G. M. (1999). Forest survey sampling designs: a history. *Journal of Forestry, 97*, 4–10.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics, 19*, 1–67.

Friedman, J. H., & Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in Medicine, 22*, 1365–1381.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *The Annals of Statistics, 28*, 337–374.

Graham, C. H., & Hijmans, R. J. (2006). A comparison of methods for mapping species ranges and species richness. *Global Ecology and Biogeography, 15*, 578–587.

Graham, J., Newman, G., Jarnevich, C., Shory, R., & Stohlgren, T. J. (2007). A global organism detection and monitoring system for non-native species. *Ecological Informatics, 2*, 177–183.

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

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

Hirzel, A. H., Hausser, J., Chessel, D., & Perrin, N. (2002). Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology, 83*, 2027–2036.

Hook, P. B., Burke, I. C., & Lauenroth, W. K. (1991). Heterogeneity of soil and plant N and C associated with individual plants and openings in north american shortgrass steppe. *Plant and Soil, 138*, 247–256.

Huang, C., Wylie, B., Yang, L., Homer, C., & Zylstra, G. (2002). Derivation of a tasselled cap transformation based on landsat 7 at-satellite reflectance. *International Journal of Remote Sensing, 23*, 1741–1748

Kumar, S., Spaulding, S. A., Stohlgren, T. J., Hermann, K. A., Schmidt, T. S., & Bahls, L. L. (2009). Potential habitat distribution for the freshwater diatom *Didymosphenia geminata* in the continental US. *Frontiers in Ecology and Environment, 7*, 415–420.

Levin, S. A. (1992). The problem of pattern and scale in ecology. *Ecology, 73*, 1943–1967.

Li, G. Y., & Weng, Q. H. (2005). Using landsat ETM plus imagery to measure population density in Indianapolis,

Indiana, USA. *Phoyogrammetric Engineering and Remote Sensing, 71*, 947–958.

Mack, R. N., Simberloff, D., Lonsdale, W. M., Evans, H., Clout, M., & Bazzaz, F. A. (2000). Biotic invasions: causes, epidemiology, global consequences, and control. *Ecological Applications, 10*, 689–710.

Marilyn, S. J., & Hart, R. H. (1994). 61 years of secondary succession on rangelands of the wyoming high-plains. *Journal of Range Management, 47*, 184–191.

Mateo-Tomas, P. & Olea, P.P. (2010). Anticipating knowledge to inform species management: predicting spatially explicit habitat suitability of a colonial vulture spreading its range. *Plos One*, 5:e12374

Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods, 5*, 434–458.

Medley, K. A. (2010). Niche shifts during the global invasion of the asian tiger mosquito, aedes albopictus skuse (culicidae), revealed by reciprocal distribution models. *Global Ecology and Biogeography, 19*, 122–133.

Menke, S. B., Holway, D. A., Fisher, R. N., & Jetz, W. (2009). Characterizing and predicting species distributions across environments and scales: Argentine ant occurrences in the eye of the beholder. *Global Ecology and Biogeography, 18*, 50–63.

Miller, J. R., Turner, M. G., Smithwick, E. A. H., Dent, C. L., & Stanley, E. H. (2004). Spatial extrapolation: the science of predicting ecological patterns and processes. *Bioscience, 54*, 310–320.

NEON (2010). National Ecological Observatory Network Inc. Available at: http://www.neoninc.org/

Parisien, M. A., & Moritz, M. A. (2009). Environmental controls on the distribution of wildfire at multiple spatial scales. *Ecological Monographs, 79*, 127–154.

Patman, J. P., & Hugh, I. H. (1961). Preliminary reports on the flora of wisconsin. No. 44. Cruciferae–mustard family. Wisconsin Academy of Science. *Arts and Letters, 50*, 17–73.

Pearce, J., & Ferrier, S. (2001). The practical value of modelling relative abundance of species for regional conservation planning: a case study. *Biological Conservation, 98*, 33–43.

Pearson, R. G., Thuiller, W., Araujo, M. B., Martinez-Meyer, E., Brotons, L., Mcclean, C., et al. (2006). Model-based uncertainty in species range prediction. *Journal of Biogeography, 33*, 1704–1711.

Penman, T. D., Pike, D. A., Webb, J. K., & Shine, R. (2010). Predicting the impact of climate change on australia's most endangered snake, hoplocephalus bungaroides. *Diversity and Distributions, 16*, 109–118.

Peterson, A. T., Papes, M., & Eaton, M. (2007). Transferability and model evaluation in ecological niche modeling: a comparison of garp and maxent. *Ecography, 30*, 550–560.

Phillips, S. J., & Elith, J. (2010). POC plots: calibrating species distribution models with presence-only data. *Ecology, 91*, 2476–2484.

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

Potts, J. M., & Elith, J. (2006). Comparing species abundance models. *Ecological Modelling, 199*, 153–163.

Randin, C. F., Dirnbock, T., Dullinger, S., Zimmermann, N. E., Zappa, M., & Guisan, A. (2006). Are niche-based species distribution models transferable in space? *Journal of Biogeography, 33*, 1689–1703.

R Development Core Team (2010). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rollins, M. G. (2009). Landfire: a nationally consistent vegetation, wildland fire, and fuel assessment. *International Journal of Wildland Fire, 18*, 235–249.

Sagarin, R. D., Gaines, S. D., & Gaylord, B. (2006). Moving beyond assumptions to understand abundance distributions across the ranges of species. *Trends in Ecology & Evolution, 21*, 524–530.

Scott, M. J., Heglund, P. J., & Morrison, M. L. (2002). *Predicting species occurrences: issues of accuracy and scale*. Washington: Island Press.

Scull, P., Franklin, J., & Chadwick, O. A. (2005). The application of classification tree analysis to soil type prediction in a desert landscape. *Ecological Modelling, 181*, 1–15.

Soil Survey Staff (1999). Soil taxonomy: a basic system of soil classification for making and interpreting soil surveys (2nd edn). Washington, DC: US Department of Agriculture Soil Conservation Service

Stohlgren, T. J., & Schnase, J. L. (2006). Risk analysis for biological hazards: what we need to know about invasive species. *Risk Analysis, 26*, 163–173.

Stohlgren, T. J., Chong, G. W., Schell, L. D., Rimar, K. A., Otsuki, Y., Lee, M., et al. (2002). Assessing vulnerability to invasion by nonnative plant species at multiple spatial scales. *Environmental Management, 29*, 566–577.

Stohlgren, T. J., Ma, P., Kumar, S., Rocca, M., Morisette, J. T., Jarnevich, C. S., et al. (2010). Ensemble habitat mapping of invasive plant species. *Risk Analysis, 30*, 224–235.

Strubbe, D., Matthysen, E., & Graham, C. H. (2010). Assessing the potential impact of invasive ring-necked parakeets *Psittacula krameri* on native nuthatches *Sitta europeae* in Belgium. *Journal of Applied Ecology, 47*, 549–557.

Thuiller, W., Brotons, L., Araujo, M. B., & Lavorel, S. (2004). Effects of restricting environmental range of data to project current and future species distributions. *Ecography, 27*, 165–172.

Vanderwal, J., Shoo, L. P., Johnson, C. N., & Williams, S. E. (2009). Abundance and the environmental niche: environmental suitability estimated from niche models predicts the upper limit of local abundance. *The American Naturalist, 174*, 282–291.

Wiens, J. A. (1989). Spatial scaling in ecology. *British Ecological Society, 3*, 385–397.

Wiens, J. A., Stralberg, D., Jongsomjit, D., Howell, C. A., & Snyder, M. A. (2009). Niches, models, and climate change: assessing the assumptions and uncertainties. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 19729–19736.

Wikum, D. A., & Shanholtzer, G. F. (1978). Application of braun-blanquet cover-abundance scale for vegetation analysis in land-development studies. *Environmental Management, 2*, 323–329.

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., et al. (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions, 14*, 763–773.