Global Change Biology

Global Change Biology (2015) **21**, 4464–4480, doi: 10.1111/gcb.13038

Projecting future expansion of invasive species: comparing and improving methodologies for species distribution modeling

KUMAR P. MAINALI[1], DAN L. WARREN[2], KUNJITHAPATHAM DHILEEPAN[3], ANDREW MCCONNACHIE[4,5], LORRAINE STRATHIE[4], GUL HASSAN[6], DEBENDRA KARKI[7], BHARAT B. SHRESTHA[8] and CAMILLE PARMESAN[9,10]

[1]*Department of Integrative Biology, mail code C0930, The University of Texas at Austin, Austin, TX 78712, USA,* [2]*Department of Biological Sciences, Bldg. E8B, Macquarie University, Sydney, NSW 2109, Australia,* [3]*Department of Agriculture and Fisheries, Ecosciences Precinct, Biosecurity Queensland, GPO Box 267, Brisbane, Qld 4001, Australia,* [4]*Agricultural Research Council-Plant Protection Research Institute, Private Bag X6006, Hilton 3245, South Africa,* [5]*Weed Research Unit, Biosecurity, NSW Department of Primary Industries, Locked Bag 6006, Orange, NSW 2800, Australia,* [6]*Department of Weed Science, NWFP Agricultural University, Peshawar 25130, Pakistan,* [7]*College of Applied Sciences Nepal, Anamnagar, Kathmandu, Nepal,* [8]*Central Department of Botany, Tribhuvan University, Kirtipur, Kathmandu, Nepal,* [9]*Marine Institute, Plymouth University, Marine Bldg. rm 305, Drakes Circus, Plymouth PL4 8AA, UK,* [10]*Department of Geological Sciences, mail code C9000, The University of Texas at Austin, Austin, TX 78712, USA*

Abstract

Modeling the distributions of species, especially of invasive species in non-native ranges, involves multiple challenges. Here, we developed some novel approaches to species distribution modeling aimed at reducing the influences of such challenges and improving the realism of projections. We estimated species–environment relationships for *Parthenium hysterophorus* L. (Asteraceae) with four modeling methods run with multiple scenarios of (i) sources of occurrences and geographically isolated background ranges for absences, (ii) approaches to drawing background (absence) points, and (iii) alternate sets of predictor variables. We further tested various quantitative metrics of model evaluation against biological insight. Model projections were very sensitive to the choice of training dataset. Model accuracy was much improved using a global dataset for model training, rather than restricting data input to the species' native range. AUC score was a poor metric for model evaluation and, if used alone, was not a useful criterion for assessing model performance. Projections away from the sampled space (*i.e.*, into areas of potential future invasion) were very different depending on the modeling methods used, raising questions about the reliability of ensemble projections. Generalized linear models gave very unrealistic projections far away from the training region. Models that efficiently fit the dominant pattern, but exclude highly local patterns in the dataset and capture interactions as they appear in data (e.g., boosted regression trees), improved generalization of the models. Biological knowledge of the species and its distribution was important in refining choices about the best set of projections. A *post hoc* test conducted on a new Parthenium dataset from Nepal validated excellent predictive performance of our 'best' model. We showed that vast stretches of currently uninvaded geographic areas on multiple continents harbor highly suitable habitats for parthenium. However, discrepancies between model predictions and parthenium invasion in Australia indicate successful management for this globally significant weed.

Keywords: AUC, boosted regression trees, generalized additive models, generalized linear models, invasive species, model evaluation, nonequilibrium distribution, *Parthenium hysterophorus*, random forests, species distribution modeling

Received 16 November 2014 and accepted 10 June 2015

Introduction

A main challenge in predicting geographic spaces likely to provide suitable habitat to an invasive species is the identification of appropriate correlates of successful vs. unsuccessful invasion (e.g., environmental variables and biotic interactions). Long-term establishment of a species in a region requires an intersection of (i) environmental conditions favorable for survivorship and reproduction, (ii) biotic interactions that are not sufficiently detrimental to cause local extinction (negative biotic interactions would include competition, allelopathy, predation, disease; lack of positive biotic interactions also have a negative impact, such as lack of pollinators), and (iii) the capacity of the species to dis-

Correspondence: Camille Parmesan, tel. +44 1752 584 993, fax +44 1752 584 955, e-mail: camille.parmesan@plymouth.ac.uk

perse to areas with favorable environmental conditions and biotic interactions (Kolar & Lodge, 2001; Guisan & Thuiller, 2005; Soberón & Peterson, 2005; Soberón, 2007) . This implies that predicting species distributions can in some cases be safely performed with environmental variables alone, especially in the absence of strong biotic interactions.

Although modeling a species' distribution is always challenging (Araújo & Guisan, 2006), an additional major challenge when modeling invasive species with correlative models is that the model is often required to extrapolate from the known environmental space (which contains species occurrence records) to an unknown environmental space (non-native geographic regions that are potential areas of future invasion). Specifically, this challenge has three components:

1. *Altered species–environment relationships in the novel vs. realized niches.* Predictions made within the range of geographic space sampled for model building (the *training region*) are reliable enough because correlations between the explanatory variables tend to remain consistent across that range (Elith & Leathwick, 2009) and so interpolation in the environmental space encompassed by the training data is likely to capture the underlying relationships. Models can be used to project into unsampled geographic spaces if the species–environment relationships, the biotic interactions, and the genetic makeup of the populations (genetic variability as well as phenotypic plasticity) are sufficiently similar between sampled and unsampled areas (Austin, 2002). However, invasive populations can have altered biotic interactions (e.g., removal from competition, parasites, or predators), differences in relative importance of environmental variables, or evolutionary changes (from either genetic drift or different selection pressures in the invaded range) (Ackerly, 2003; Lavergne & Molofsky, 2007; Pearman *et al*., 2008; Duncan *et al*., 2009).
2. *Extrapolation of the models beyond the domain of parameter calibration.* Predicting beyond the domain over which parameters are calibrated can be risky because of lack of observations for model calibration and evaluation (Elith & Leathwick, 2009; Zurell *et al*., 2012). Many studies have found that the climatic space occupied by invasive species in their introduced ranges is often broader than that in their native ranges (Fitzpatrick *et al*., 2007; Loo *et al*., 2007; Kearney *et al*., 2008). Such a discrepancy in climatic space can result from the differences between native and introduced ranges discussed above, but the discrepancy can also result from the fundamental niche not being fully realized in native ranges because of (i) dispersal constraints and/or biotic interactions preventing establishment in some climatically suitable areas (Araújo & Peterson, 2012) and (ii) the geographic area historically inhabited by the species not covering the entire domain of multivariate climatic space that could support a population (Mandle *et al*., 2010). Therefore, species distribution models generated within native ranges may represent only part of the fundamental niche (Soberón & Peterson, 2005).
3. *Nonequilibrium distribution in invasive ranges.* When occurrence records are available from invaded ranges, pairing these occurrences with background samples is challenging because invaded ranges in which the species may still be expanding in extent or abundance represent a case of nonequilibrium distribution (Thuiller *et al*., 2005; Rodda *et al*., 2011). Even though species within their native ranges often occupy fewer areas than are suitable (*i.e.,* their realized niche is smaller than their fundamental niche), the plant in its native range occurs at some level of equilibrium distribution across all suitable pixels, whereas the plant in regions it is actively invading is, by definition, *not* in spatial equilibrium. Therefore, unoccupied spaces in invaded ranges have higher chances of harboring environmentally suitable habitat than in native ranges, simply due to insufficient time having passed for the species to occupy the full extent of suitable habitat that it is capable of occupying.

Studies have attempted to address these challenges. First, when the observed climatic niche differs between native and non-native ranges (Broennimann *et al*., 2007), models calibrated in one geographic region can underperform in new geographic spaces (Fitzpatrick *et al*., 2007; Beaumont *et al*., 2009). This challenge of limited model transferability across space can be dealt with by inclusion of both native and non-native ranges in model training, which improved projection in invaded ranges in some studies (Mau-Crimmins *et al*., 2006; Broennimann & Guisan, 2008; Beaumont *et al*., 2009).

Second, as Monahan (2009) showed with a mechanistic niche model, the challenge of nonequilibrium distributions can arise because the realized niche can be smaller than the fundamental niche due to dispersal constraints, biotic interactions, and other reasons. These conditions, in addition to the issues imposed by ongoing range expansion, make invasive distributions far from representative of a species' potential equilibrium distribution. While the challenges of nonequilibrium distribution cannot be eliminated entirely, model reliability can be improved with the use of expert opinion (Murray *et al*., 2009).

We selected one species as a test case to examine these complex issues. We modeled the present and

potential future distribution of the invasive plant, *Parthenium hysterophorus* L. (Asteraceae; parthenium). *Parthenium hysterophorus* is a globally significant weed that has invaded Asia, Africa, and Australia (>30 countries in total) (Adkins & Shabbir, 2014). From its pattern and degree of spread, parthenium appears to be primarily climatically limited. There are presently no known strong biotic interactions that restrict the distribution of parthenium at broad spatial scales, and given the near-global distribution of the species (Fig. 1), it is likely that such interactions are of minor importance to its establishment. Therefore, parthenium represents an excellent opportunity for exploration of the robustness of differing methodologies within the broad realm of environmental species distribution modeling (SDM), with the aim of developing 'best practices' for modeling spread of invasives in general, and specifically estimating areas at high risk of future invasion by parthenium.

Here, we use *P. hysterophorus* as a case study to develop novel approaches to correlative SDM aimed at reducing the influences of these challenges and improving the realism of projections. First, we propose a new approach designed to (i) improve model transferability across space (i.e., from training region into new geographic spaces) and (ii) reduce the chance of sampling false absences of species in a nonequilibrium state of distribution. This approach uses occurrences from all regions but obtains background (absence) points only from native ranges. We then present approaches for modeling the invasive species at a global scale; specifically, we quantitatively compare the effect of the following in predicting the species distribution in native ranges, invaded ranges, and potential areas for future spread: (i) sources of occurrences and background ranges, (ii) approaches to drawing background points, and (iii) alternate sets of predictor variables. We also compare the accuracy of different modeling methods in projecting occurrences far away from the training region and relate these results to AUC scores within the training region.

Materials and methods

Distribution, invasion history, and biology

Parthenium (*Parthenium hysterophorus* L., Asteraceae), a native of Central America, Mexico, and southeastern USA, is a weed of global significance (Navie *et al*., 1996). The plant was first identified in non-native ranges as a weed in Queensland, Australia, in 1955 (Auld *et al*., 1982–83) and then India in 1956 (Rao, 1956). Since the 1950s, parthenium has spread to most humid/subhumid tropical and subtropical areas of the world, from sea level to 2700 m (Dhileepan & McFadyen, 2012). Genetic analysis suggests that parthenium genotypes found in Australia, India, and Africa possibly originated from southern Texas, USA (Graham & Lang, 1998).

Parthenium is an annual herb with a deeply penetrating taproot and an erect shoot. With good rainfall and warm temperature, parthenium has the ability to germinate and establish at any time of the year. Parthenium is a prolific seed producer; a mature plant can produce more than 150 000 seeds in its lifetime (Dhileepan, 2012). The seed is spread by animals, wind, water, vehicles, agricultural and road construction machinery, fodder, and seed lots (Auld *et al*., 1982–83; Navie *et al*., 1996), as well as other human activities (e.g., parthenium flowers in bouquets, green parthenium plants as packing materials, and parthenium weed as green manure). Buried seeds persist and remain viable in soil for reasonably long periods, with nearly 50% of the seed bank viable up to 6 years (Navie *et al*., 1998). In the invaded ranges, parthenium negatively affects crops, rangeland productivity, native biodiversity, and the health of humans and animals (reviewed in Dhileepan, 2009).

Resolution and extent of study areas

Because of parthenium's unusual success in spreading to all continents except Europe, our study modeled its future distribution on a global scale. We performed the modeling at 2.5-arc-min resolution. We excluded Antarctica from analyses, as very little of that continent is suitable for plant life.

Occurrence records

We obtained occurrence records from freely available databases, published personal records, and primary data collected for this study (see Appendix S1, Table S1). We eliminated points with a spatial uncertainty greater than 1 min, yielding 3989 points, averaging 1.7 occurrences per grid cell. However, there was a marked variation in density across the sources. For instance, one source (coauthor DK) had >37 occurrences per grid cell (859 records in 23 grid cells). DK confirmed that he performed an exhaustive survey of the plants in several patches of the 23 grid cells. Based on our field observation, the surrounding habitat is similarly suitable for the plant but we have an order of magnitude fewer points from it. Therefore, to minimize the effect of sampling bias (e.g., Elith *et al.*, 2010), we eliminated all but one point per grid cell, yielding 2322 points for analyses. This approach, which eliminates all but one presences within *ca* 5 km by 5 km area, is similar to the spatial filtering of occurrences by Boria *et al.* (2014) where they eliminated presences within 10 km of a selected occurrence record and yielded better models as a result of reduced sampling bias and overfitting.

Assessing the role of roads

Roads have been shown to be associated with spread of invasive plants (Tyser & Worley, 1992; Parendes & Jones, 2000). In invaded areas, parthenium records also tend to occur near roads. We tested the role of road for its facilitation effect, as a

conduit of propagule dispersal and as a driver of spatial sampling bias (see Appendix S2-A for details).

Choice of background points

Eight minimum area convex polygons were created around concentrated regions of occurrences (Fig. 1). We chose this method based on previous research showing that models built using a geographic background much larger than the core area of the species' distribution can result in poor model performance: Acevedo *et al.* (2012) found that increasing the geographic extent of the background results in higher discriminatory power of the model within the background, with an increase in AUC. However, when the same models were evaluated with records from the core area of the distribution, a negative relationship was observed between geographic extent and AUC, reducing the reliability of the models in projecting core area of distribution. To reduce the chances of models with artificially inflated AUC but with little real-world relevance when projected, we limited the background ranges to the most concentrated areas of occurrence. This resulted in 3.4% of the presences falling outside of the background regions but effectively reduced the background regions to about one-third of the area of convex polygons created separately within each continent encompassing all presences of the continent. However, those 3.4% of presences that fell outside of the selected background areas were retained in the list of presences, making use of all the occurrence records in the study.

Background points were not drawn from the space within 10 km of recorded presences. To keep prevalence (the proportion of sites with presences, or number of presences/number of both types of points) constant between regions, we matched the number of background points to the presences within each region. Barbet-Massin *et al.* (2012) show that regression models (GAM and GLM) do not substantially improve with an increase in the number of background points to those typically suggested for MaxEnt (e.g., 10 000), and classification models actually get worse with larger numbers of such points; they further suggest using same number of presences and background points for RF and BRT, providing support to our fairly large dataset (2322 points of each type) and the design of equal number of two types of points. A random draw of background points assumes that the grid cells are of equal size because each grid cell has equal chance of being selected. In reality, grid cells further away from the equator are progressively smaller because of the Earth's curvature. Background samples therefore need to be drawn taking into account cell sizes if the latitudinal gradient in the range is nontrivial (>200 m; Elith *et al.*, 2011), which is the case in this study. We therefore undertook weighted sampling such that grid cells were sampled in proportion to their geographic area. To estimate the effect of roads on sampling bias, we drew one set of background points using only cell area as the weight/bias (AreaBias) and a second set weighted using both cell area and linear distance to roads (AreaRoadBias).

Predictor variables

We obtained raster layers for 19 climatic variables and altitude at 2.5-arc-min resolution from WorldClim version 1.4 (Hijmans *et al.*, 2005, www.worldclim.org). This set of climatic variables (Appendix S1, Table S2) was supplemented with other variables that are likely to affect parthenium: soil moisture, percent canopy cover, human population density, and distance to the nearest road (linear and square root, Appendix S2-A, Fig. S1). See Appendix S1 and Table S3 for additional information about variables.

Species distribution models

We used two regression based models, that is, generalized linear models (GLM) and generalized additive models (GAM), and two decision tree based methods, that is, random

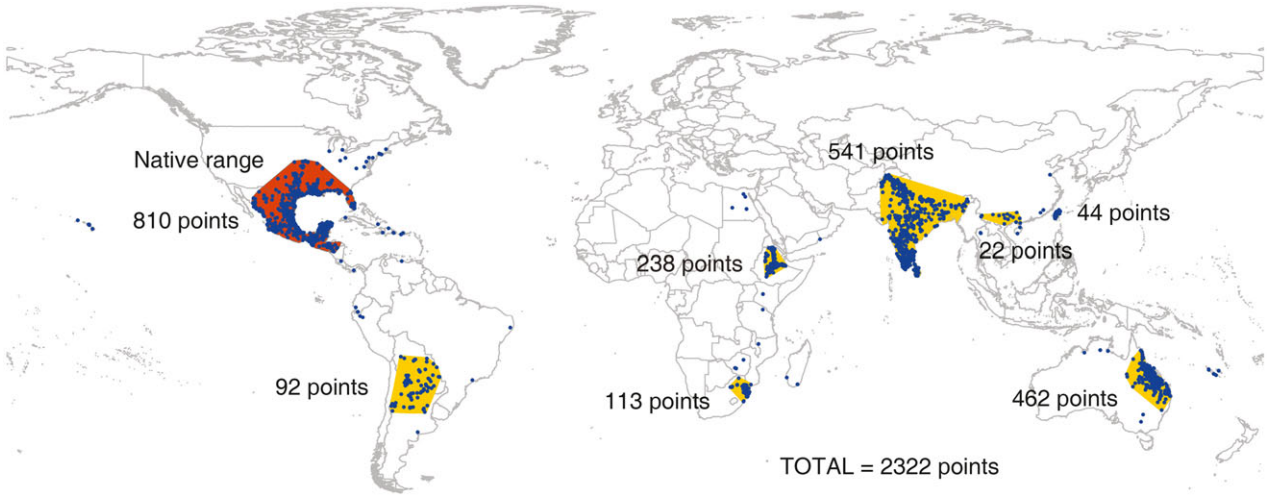

Fig. 1 Occurrence records (blue solid circles) and background regions (yellow polygons, orange in native range). Numbers displayed next to each of the eight polygons represent the number of presence points drawn, and is equal to the number of background points drawn.

forests (RF) and boosted regression trees (BRT). These four modeling methods have been, in general, shown to perform well in SDMs (Araújo *et al.*, 2005; Elith *et al.*, 2006; Pearson *et al.*, 2006; Elith & Graham, 2009) but each has their own strengths, biases, and weaknesses. Modeling distributions of invasive species has been performed with high accuracy using BRT, RF, and GAM (Cutler *et al.*, 2007; Broennimann & Guisan, 2008; Elith *et al.*, 2010). Because the same sets of data were used for training and testing all methods, the only differences between models being compared were the modeling methods themselves. This allowed us to isolate the effects of the methods when comparing models. When we conducted the analysis in the BIOMOD package of R, MaxEnt (Phillips *et al.*, 2006) – one of the most popular modeling algorithms in SDM – was not available in the package. Running MaxEnt models in its stand-alone software presented important problems that we could not resolve: We applied two types of biases while drawing background points which were drawn in fixed number from each of eight regions of the world. Then, fivefold partitioning of the presences and background points was performed for each continent separately. This was not possible with the MaxEnt stand-alone software, so MaxEnt was omitted from this study.

Overfitting and predictive performance

An excessively complex model has very high fit to the training data because its excess parameters (relative to the number of observations) explain random error in the data. This can obscure the true underlying relationship between variables and therefore yields a model with poor predictive performance. We used two approaches to control overfitting. The Akaike information criterion (AIC) was used for GLMs. Cross-validation was used for GAMs, RF, and BRTs.

Various novel combinations of background sampling method, pairing of presences to background points, and choices of predictor variables

We performed nonmetric multidimensional scaling (NMDS) of 23 environmental variables used in SDM and plotted occurrences in the ordination plot; principal components analysis (PCA) was not suitable for extracting components because of highly nonlinear relationships between the predictor variables. We developed three methods for selecting data points to train models: (i) presence points from the world and background points from various polygons in the world (PWBW), (ii) both presence and background points from native ranges (PNBN), and (iii) presence points from the world and background points from the native range (PWBN) (Fig. 2).

The background points in each of the three point sources were drawn using two biases: (i) cell area (AreaBias) such that background points were more likely to be drawn from bigger cells and (ii) both cell area and proximity to road (AreaRoad-

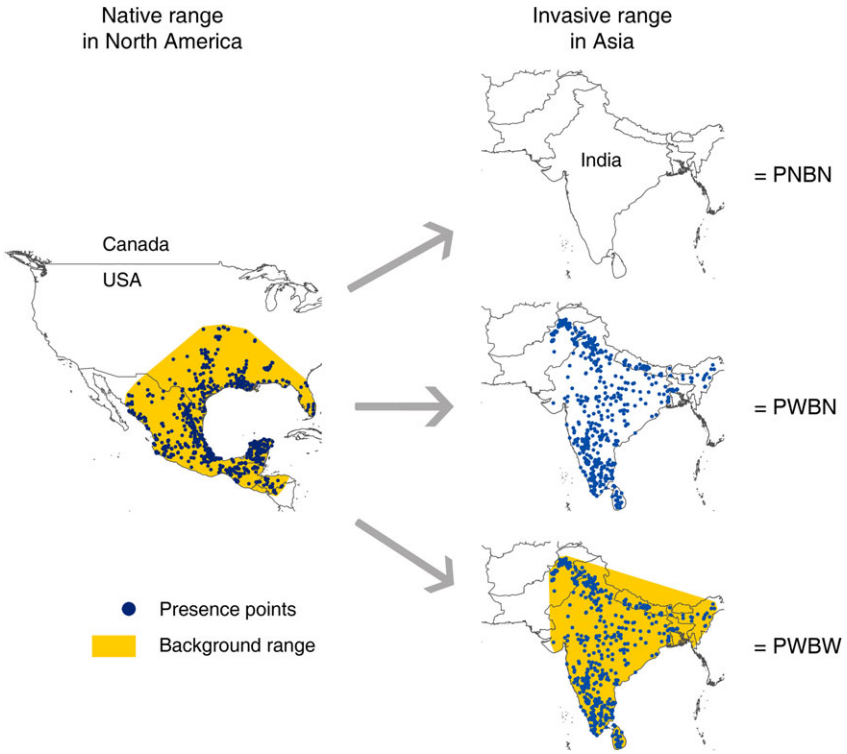

Fig. 2 Visual description of the three methods for selecting data points to train models. PNBN = both presence and background points from native ranges; PWBN = presence points from the world and background points from the native range; PWBW = presence points from the world and background points from various polygons in the world.

Bias) such that, on the top of size, cells nearer to roads are more likely to be selected than those further away. We created two sets of explanatory variables: (i) WorldClim, soil moisture, percent canopy cover, human population density, (ii) all variables in the first set plus proximity to road (both linear and square root) (Appendix S1, Table S2). We performed the study with a fully crossed design of these three factors; the design gave us a set of 12 combinations (hereafter 'scenarios') of point source, bias in drawing background, and sets of explanatory variables (Table 1).

Evaluation indices

We evaluated models with the following metrics: area under the receiver operating characteristic curve (AUC), sensitivity, specificity, Cohen's kappa, and the true skill statistic (TSS). AUC scores are easy to interpret and have been widely used in comparing species distribution models, but have recently been criticized for several reasons (Allouche *et al*., 2006; Lobo *et al*., 2008). We dealt with several of these criticisms in the following ways: (i) An ROC plot, and therefore the AUC score, does not provide information about the distribution of model errors in geographic space. We dealt with this criticism by computing AUC scores for each continent separately, as well as for the entire sampling extent and the world; (ii) AUC scores can easily be inflated by increasing the geographic extent for drawing background points. To deal with this criticism, we set geographic backgrounds in eight convex polygons enclosing dense masses of occurrences, leaving out isolated points, and reducing the background area dramatically. We then used the same set of points for all the models within each of the three levels of the factor 'point source' (Table 1). The three levels of 'point source' were intended to be different in their geographic extent of sampling ranges, so that we could test the effect of point sources in models; (iii) Obtaining random background points from sites that are not confirmed for species' absences inflates the chances of false absences. This is unlikely in our study to cause differences among methods, as the same set of presence and background points were used for each modeling method. Finally, the potential effect of prevalence was minimized using the same number of presence and background points.

In contrast to AUC, the benefit of using Cohen's kappa is that it corrects for the model fit expected by chance (Allouche *et al*., 2006). However, Cohen's kappa is sensitive to prevalence. Allouche *et al.* (2006) therefore recommend using TSS for model evaluation.

Traditional vs. region-specific model evaluation

AUC and other evaluation metrics computed on independent data provide estimates of model generalization and predictive power, but only within the range of sampling. The ability of a model to predict outside the training region cannot be estimated with the conventional approach of computing AUC on independent data withheld from model construction. To deal with this problem, we computed AUC and other evaluation scores for every model using presences and background points from each continent separately, with the exception of Europe for which there were no occurrence records. All AUC values reported in this study were computed in this way. We compared this AUC with the traditional AUC (computed on independent data from the training region) in Fig. 4. Our approach of computing AUC not only provided an index for comparing models' predictive capacity outside its range (*i.e.*, transferability), but also allowed us to determine the best model for projecting in each continent. Given the fact that continents have very different environmental spaces of presences (Fig. 3), it is likely there is not a single best model for predicting every continent.

Analysis and computation

The main work of species distribution modeling was performed with the package BIOMOD 1.1-7.02 (Thuiller, 2003; Thuiller *et al*., 2009) installed in R 2.14.0 (The R Project for Statistical Computing) on the Lonestar supercomputer at the Texas Advanced Computing Center. For each of the 12 scenarios (Table 1), we performed 100 independent modeling replicates. Each replicate is the average of 25 iterations resulting from sets of cross-validation points: For each random set of points (all presences, randomly drawn background points), we performed fivefold cross-validation of the models, using four groups as training sets and the fifth as a testing set. We thus obtained five sets of training presences, which we crossed

Table 1 Complete factorial design of the study. The four factors result in a total of 48 combinations of levels. Each combination had 100 independent projections of global modeling (each independent projection being an average of 25 iterations resulting from fivefold partitioning of cross-validation sets from each random draw of background crossed with the same of presences), yielding a total of 4800 independent projections for the world. (See Fig. 2 for 'point source' abbreviations)

	Factors			
	Point source	Bias used in background draw	Explanatory variable sets	Model
Levels	PWBW PWBN PNBN	Grid cell area (AreaBias) Grid cell area and proximity to road (AreaRoadBias)	All variables including Road (Road) All variables except Road (NoRoad)	Generalized linear models (GLM) Generalized additive models (GAM) Random forests (RF) Boosted regression trees (BRT)

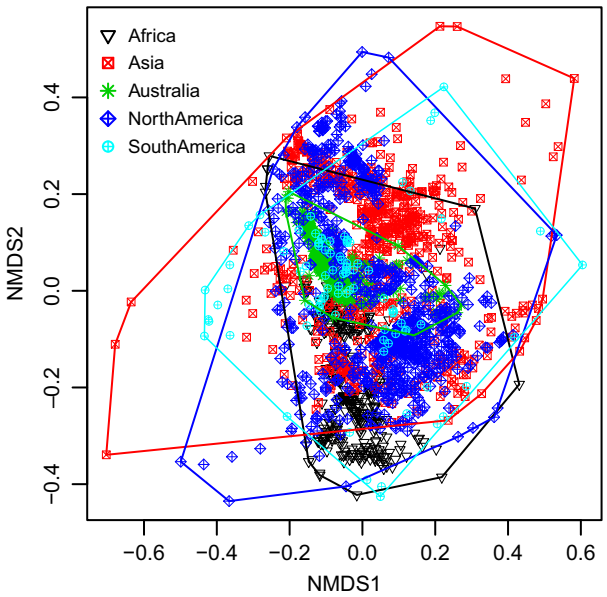

Fig. 3 Distribution of presences from different continents in the first two axes of a nonmetric multidimensional scaling (NMDS) of 23 environmental predictors (#1–23 in Appendix S1, Table S2). To test whether presences from different continents occupy similar ecological niche, we conducted Welch's ANOVA and Levene's test for homogeneity of variance (as in Mandle *et al.*, 2010) and two other tests. The continents are significantly different along each of first two NMDS axes. For the first NMDS axis, Bartlett test of homogeneity of variances: *K*-squared = 430.4321, df = 4, *P*-value < 2.2e-16 (Levene's test yielding highly significant difference also); one-way analysis of means with Welch's correction: *F* = 129.2033, num df = 4.000, denom df = 530.801, *P*-value < 2.2e-16 (Kruskal–Wallis rank sum test yielding highly significant difference also). For the second NMDS axis, Bartlett test: *K*-squared = 662.1226, df = 4, *P*-value < 2.2e-16 (similar results by Levene's test); Welch's ANOVA: *F* = 354.8588, num df = 4.000, denom df = 514.089, *P*-value < 2.2e-16 (similar results by Kruskal–Wallis test). Tukey's multiple comparisons of means were significant at 0.05 level for every pairwise comparison of continents in at least one axis of the plot.

with five sets of training background points, yielding a total of 25 projections. As the five sets of occurrence points were not truly independent of each other (once a set of points are divided into five groups and the first set of training and testing points are created, all the other sets of training and testing points can be predicted), the resulting 25 projections were averaged to obtain one independent projection. In total, we generated 12 scenarios * 4 SDMs * 100 independent replicates = 4800 projections. The BIOMOD settings included the following: polynomial terms and stepwise procedures using AIC criteria for GLM, maximum number of trees to be 5000 for BRT, and three degrees of smoothing in spline functions for GAM. Analysis of BIOMOD output and plotting was performed in the following packages installed to R 2.15.1: gridExtra, matrixStats, plyr, PresenceAbsence, R.methodsS3, Sciplot, sperrorest, TeachingDemos, and AUC.

Incorporation of expert opinion

The eight regions (Fig. 1) where models were trained/tested comprise only 7.2% of all grid cells where models were projected. Outside of these polygons, the relevance of the evaluation metric can be questionable (see 'Introduction' for three main reasons). Therefore, we needed some basis to evaluate the models outside of those polygons (93% of the grid cells). For determining the best model for each continent, we supplemented AUC scores (useful for evaluating the models within training/testing ranges) with expert opinion (useful for evaluating the models outside of model training/testing ranges). Expert opinion did not replace or undermine AUC scores but rather added to the model selection process. For incorporating expert opinion in the model selection process, the first author (KM) presented 48 projections of the world (see Table 1 for the combinations of factors) to three experts on parthenium (coauthors KD, AM, LS), each of whom has spent extensive time studying Parthenium under both field and laboratory conditions. Each expert was interviewed separately as to how the model projections matched up to their own experiences for the region they knew. The three experts have conducted extensive field work on many aspects of parthenium ecology and management, including extensive distribution surveys as well as studies of seed banks, natural herbivores, and management options (e.g., introduced biocontrol agents and postrelease evaluation) across the entire current range in 15 countries (South Africa, Mozambique, Swaziland, Ethiopia, Kenya, Tanzania, Bolivia, Brazil, Paraguay, Madagascar, Venezuela, Australia, Argentina, India, and Sri Lanka). Each expert offered their opinion about the realism of the model projections based upon over a decade-long field experience with parthenium management in Africa, Asia, or Australia, and upon cumulative understanding about the requirements and tolerances of this plant across a range of climatic and environmental conditions present in suitable habitats across the world. Each expert recommended the best model for each continent after examining different parts of the continent for the mismatch between projected and expected habitat suitability. Extended details about the method are provided in Appendix S2-C.

Results

Continental differences in the multivariate environmental space of presence points

In the first two axes of a nonmetric multidimensional scaling (NMDS) plot of 23 predictor variables, clusters of occurrence records from various continents had a markedly different extent, central tendency, and dispersion ($P \ll 0.0001$, Fig. 3). This indicates that the environmental space of presence points in various invaded regions is different from each other and also is different

from that of the native range (Tukey's multiple comparisons of means, $P < 0.05$).

Significant effects of methodologies and choices used to construct models

The full factorial design of this study allowed us to tease apart the effects of variations of each of the four factors on modeling performance when the effects of the other factors were held constant (Table 1). We calculated AUC, sensitivity, specificity, kappa, and TSS as a performance measure of modeling methods. A four-way analysis of variance showed that all four factors – point sources, method used to draw background (absence) points, choice of explanatory variables, and choice of SDM – had significant effects on each of the five measures of model performance ($P < 0.0001$; AUC results in Appendix S1, Tables S4).

Spatial structure in occurrence points and road as a predictor

Our factorial design showed that the suspected road–weed association was not strong (see Appendix S2-A for details). When road was included as an explanatory variable, AUC improved by 0.03–0.04 but the model yielded a biologically unrealistic projection map (Appendix S2, Fig. S3) which contradicted ground surveys; three coauthors of this study (KD, AM, and LS), all with extensive experience in parthenium management throughout its invaded ranges, concluded that there was an overly dominant effect of road, with a predicted distribution unrealistically restricted to be near roads. This could result from a simple sampling bias, in which occurrences are more likely to be detected near roads due to a bias in the frequency of visits by observers. We minimized this possible source of sampling bias (for more efficient SDMs as in Syfert *et al*., 2013) by drawing more background points near roads. However, this approach (Road as a bias) did not yield significantly different AUC scores (Appendix S2, Fig. S2), suggesting that the suspected road–weed association does not exist or that spatial correlation between roads and the environmental variables used in this study is not sufficient to contribute significant bias to models. On the other hand, if the association was strong and the weighting factor (i.e., the linear distance) we used did not completely cancel out the sampling bias in presences, then road could still appear as a significant predictor without showing any bias effect in sampling. With these results, we cannot conclusively determine whether an association exists between roads and probability of presence, or if it existed, whether it resulted from sampling bias or facilitation of establishment and growth by roads. If the correlation between habitat suitability and distance to road is real, then the 'road' model would have limited application in global modeling of potential invasive spread. Therefore, for the rest of the analyses except Fig. 6, we dropped AreaRoadBias and road as a predictor.

Continent-wise prediction and predictability inside vs. outside the training region

This left only two factors: choice of training regions from which to draw point sources and choice of SDM. Models built with the three point sources (PWBW, PWBN, and PNBN) had dramatic differences in predictive ability. Obtaining both presences and background points from all regions of the world (PWBW) gave models with substantially higher predictive power on a global scale than models that were built with other combinations of points (PWBN, PNBN) (Fig. 4a). The predictive power of the models in non-native ranges worsened with the use of points from only the native range (either only background points or both background and presences). For Asia, Africa, and Australia, the AUC for PWBW was higher than that for other point sources by 0.12-0.26, and by 0.035-0.071 for South America. However, prediction accuracy within the native range (North America) was maximized by having both presence and background points from only native areas (PNBN), the difference with the other point sources being only 0.014-0.018. For the whole world, PWBW had an AUC that was 0.11 higher than the second best model (PNBN) (Fig. 4a, column 'World'). We therefore chose PWBW as the best combination of source and background points.

The AUC scores reported so far were the ones computed by predicting points from various continents irrespective of whether or not the continent contributed points to model construction. This AUC (e.g., AUC_{world} for all continents together) was not the same as the AUC computed by predicting an independent dataset from the training region ($AUC_{training\ region}$), something used traditionally for model comparisons. The dashed box in Fig. 4a shows that $AUC_{training\ region}$ (column 'Training region') was much higher than AUC_{world} (column 'World') for PWBN (0.928–0.654 = 0.274) and PNBN (0.841–0.712 = 0.129). Not surprisingly, for PWBW, the two AUCs were identical because the range of background sampling and presences fell in all continents.

Comparing models

The scenario of factors chosen as 'best' performing (PWBW with AreaBias, NoRoad) was applied to all four SDMs: GLM, GAM, RF, and BRT (Fig. 4b). RF

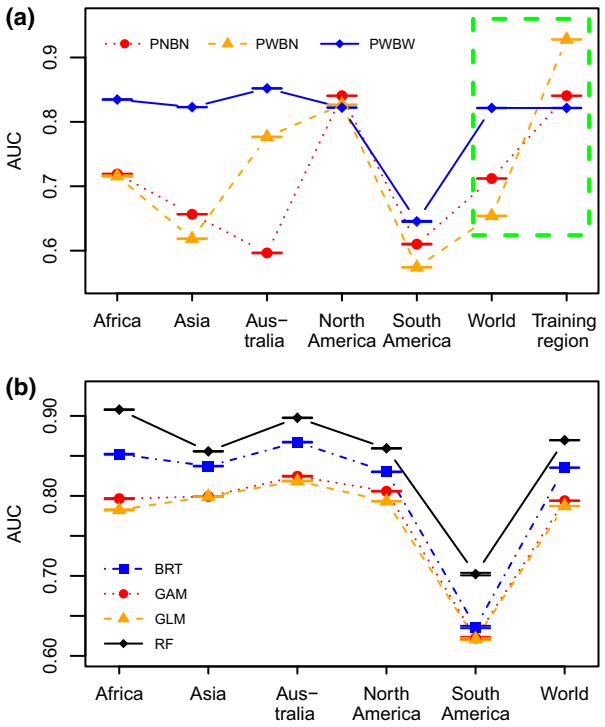

Fig. 4 AUC (±1SE) computed for different regions of the world (for comparison, kappa and TSS have similar pattern; see Appendix S3, Figs. S4, S6). (a) Point sources compared with two types of AUC score; all models collapsed. Models were trained on the 80% points of the entire dataset of each point source and tested on the held-out dataset from the same point source. AUC score computed that way is reported on column 'Training region' inside dashed box. The models were then tested for each continent separately (using presences and background points from the continent) ensuring the points used for testing were not used in model training. Weighted average of all continents (contingent upon number of points) is given in column 'World.' Within each of the seven region/continent, all pairwise differences among three point sources were significant at 0.0001 level. Dashed box shows how AUC score computed on training region is much higher than the one computed for the world. This and all the subsequent figures except Fig. 6 report result for AreaBias and NoRoad. (b) Models compared for the point source PWBW (AreaBias, NoRoad). All pairwise differences between models within a continent/region are significant at 0.05 level except the following: Asia: GAM vs. GLM, South America: GAM vs. GLM.

scored the highest AUC on every continent, with BRT second. Kappa and TSS indices followed similar patterns to AUC (Appendix S3, Fig. S4). In global comparisons, the AUC scores were as follows: RF – 0.87, BRT – 0.835, GAM – 0.794, and GLM – 0.787 (Appendix S1, Table S5). Based both on evaluation metrics and biological insight about distribution and ecophysiology of the plant (see 'Discussion' and Appendix S2-C), for our 'best' models, we chose GAM for projecting in Africa, Australia, and New Zealand, and BRT for the rest of the world (Fig. 5, and Appendix S3, Fig. S5).

Incongruence among levels of factors

The total variance of all projections for a grid cell showed a decreasing trend with increase in habitat suitability (Fig. 6a). Worldwide, most grid cells were unsuitable for parthenium. We partitioned the total variance in estimated suitability into the percentage of variance contributed by each factor. When all the grid cells were considered together, >99% of variance in suitability predictions was contributed by modeling method, point source, and choice of explanatory variables. Choice of bias and replicates of presence and background points in total accounted for <1% of the total variance (Fig. 6a, pie chart). The partitioned variances plotted against habitat suitability (Fig. 6b) exhibited a number of trends: Variation contributed by point sources decreased and variation as an effect of SDM increased with habitat suitability. For habitat suitability estimates of below 0.68, more variation was caused by point sources than by choice of SDM. For higher habitat suitability scores, differences among SDMs were responsible for more of the variance among outputs. Explanatory variable sets, bias, and background point replicates all exhibited a unimodal relationship of variation against habitat suitability, with the variation explained by each of them being highest around a habitat suitability of 0.5.

Evaluation indices

We calculated commonly used (AUC, sensitivity, specificity) and less commonly used (kappa, TSS) model evaluation indices. Our AUC scores had a very tight and linear relationship with both kappa and TSS (r = 0.85–0.89 for four SDMs, Appendix S3, Fig. S6). SDMs were given the same set of presence and background points, keeping the prevalence at 0.5. This resulted in kappa and TSS scores being identical (kappa-TS r = 1.0 for all SDMs, Appendix S3, Fig. S6), because in estimating the predictive accuracy of models, the dependence of kappa statistic on prevalence is corrected by TSS (Allouche *et al.*, 2006).

Final evaluation using expert opinion

All three experts (co-authors KD, AM and LS) came to similar conclusions about roads not being very useful as an explanatory variable for their region of expertise (discussed above). For our final choice of point source (PWBW), the recommended models were as follows

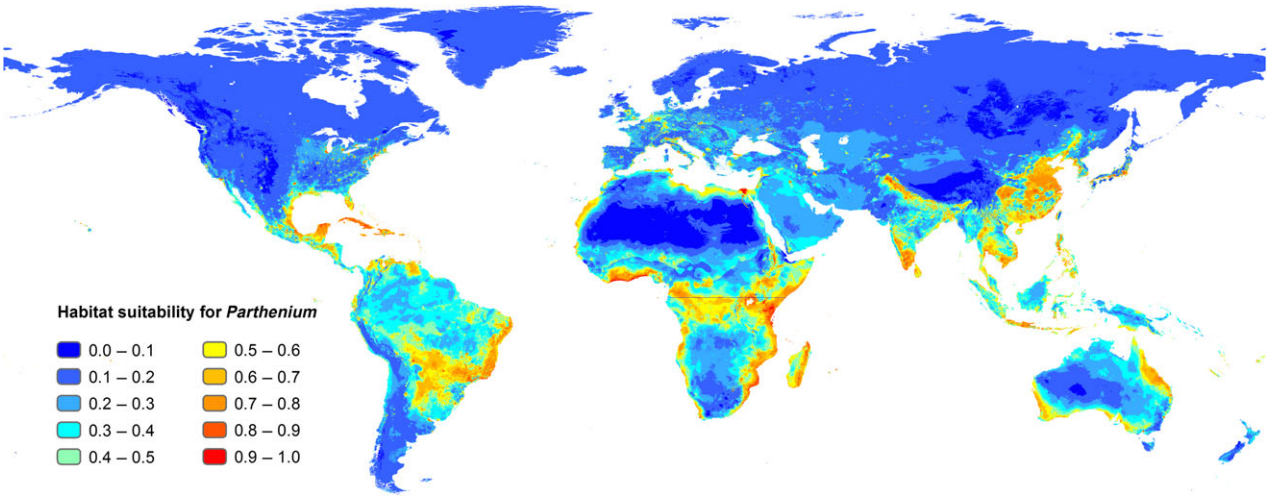

Fig. 5 Prediction of habitat suitability for the world; generalized additive models (GAM) used for Africa, Australia, and New Zealand, and boosted regression trees (BRT) used for the rest of the world. Occurrences and background points in equal number were obtained from each of the five continents (PWBW, see Fig. 1). Background points were obtained without considering proximity of grid cells to road; explanatory variables included 23 predictors but not road.

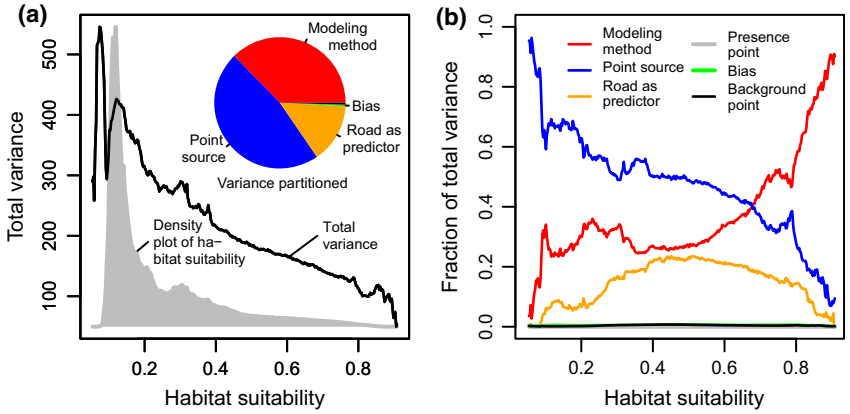

Fig. 6 Variance in 4800 independent predictions. (a) Total variance trend against habitat suitability, density plot of habitat suitability, and variance partitioned to the factors (pie chart) that contributed to it in the entire projected area (modeling method: 39.2%, point source: 47.6%, set of explanatory variables: 12.5%, bias: 0.35%, background point replicates: 0.037%, present point replicates: <0.002%); (b) rescaled variance partitioned to predictors. Variance partitioning in both plots included habitat suitability as the prediction of BRT models. The total variance in *every* grid cell was partitioned to factors and expressed as fraction for pie chart and Fig. 6b. Type I analysis of variance performed. (Note: The variance partitioned to various factors is not the fraction of the total variation in distribution explained by the factor.)

(with number of experts voting for the models in parentheses): Asia and South America – BRT (3); Australia – GAM (3); North America – BRT (2) and RF (1) with the expert voting for RF saying BRT only slightly worse than RF; and Africa – GAM (2) and BRT (1). We therefore chose BRT for Asia, North America, and South America, and GAM for Australia and Africa (Fig. 5).

Discussion

'Essentially, all models are wrong, but some are useful' (Box & Draper, 1987). SDMs in practice often use data that violate key assumptions of the models (Pearson & Dawson, 2003; Jeschke & Strayer, 2008). Specifically, it is assumed that (i) a species distribution is not affected by biotic interactions or is affected in the same way across the entire distribution, (ii) genetics and plasticity remain constant across the entire range of the distribution, and (iii) there is no dispersal constraint, allowing species to occupy all spaces with suitable climate and be absent elsewhere. Various remedies to improve the realism of SDMs have been proposed by previous studies (Broennimann & Guisan, 2008; Jiménez-Valverde *et al.*, 2011; Rodda *et al.*, 2011). Here, we demonstrated

that SDMs can be greatly improved through biological insight guiding careful selection of SDM methods, of background regions used for building the models, and of choice of predictor variables.

On top of the many challenges that always accompany SDM (Araújo & Guisan, 2006; Thuiller *et al*., 2008), modeling invasive species requires dealing with nonequilibrium distributions and often differences in climatic space occupied by the species in native and invaded ranges. We found that projections away from the sampled space were very different with different modeling methods, raising questions about the reliability of ensemble projections that average results from many different outputs. Further, traditional model evaluation indices (AUC, kappa, etc.) need careful computation and interpretation complemented with insight about the biology and distribution of the species. Biological insight becomes even more important when the projection range is much broader than the sampled geographic space.

In addition, we have demonstrated that it is also important to use model evaluation metrics computed with independent points drawn from the projected ranges, rather than from training regions. This is, as of yet, a rare practice in SDM.

To the best of our knowledge, our study is the first to quantitatively compare the effect of decoupling presences from background ranges. However, our results also demonstrated that a decoupling approach does not necessarily lead to a better model. A frequently reported challenge of SDMs is that background ranges (where the species is absent) are much larger than the range of presences, a situation that artificially inflates AUC scores. One of our choices for points, PWBN, was opposite to most other studies in that the background range was much smaller than the range of presences. We chose to examine this combination of presences and absences based on the logic that, while an invasion is still in progress (as is the case for parthenium), the invaded range will contain substantially more 'false absences' than the native range, simply because the plant has yet to invade all suitable habitat that it will eventually be able to occupy. While the biological justification for this choice of presence and background points seems sound, the statistical problems that emerged by inferring a model in this fashion resulted in models that were not particularly trustworthy. Models trained with PWBN were unreliable, predicting suitable habitats in Greenland and northern Canada where this tropical/subtropical species not only currently does not exist, but, according to our three parthenium experts, is not expected to ever be able to exist. In spite of this lack of biological realism, these same models secured the highest AUC score when evaluated with independent data from the training region (Fig. 4a, dashed box; discussed below).

Improving model performance

By approaching the global modeling of parthenium via 12 scenarios that explore the effects of geographic training region (sources of points), possible sources of sampling bias, and possible effects of roads on model outputs, we found that no single evaluation criterion was adequate for choosing the 'best' set of approaches. We found the most important areas to consider could be grouped into three themes: Choices made concerning appropriate use of model evaluation metrics, the model training region, and choice of SDM. We explore these in more detail below.

Evaluation metrics. We found that AUC scores can be very misleading if used as sole criteria for choosing a model, supporting the few previous studies that have explored this (Allouche *et al*., 2006; Lobo *et al*., 2008). Biological knowledge of the species and its distribution was important in refining choices about the best set of predictions (Murray *et al*., 2009), especially when the geographic range of predictions is much broader than the training region of the model, as is true for most invasive species.

We hypothesized that PWBN would give the best model because it would have two advantages over other point sources: (i) Occurrence points outside of the native ranges were expected to either expand the niche or more completely characterize the historic niche, and (ii) background points taken only from within the native range would be less likely to fall on suitable, but currently unoccupied habitats. AUC computed on the withheld data from sampling ranges (from the same range that provides model building points) was very high with an average AUC of four SDMs of 0.93 (see column 'Training region' in the dashed box, Fig. 4a). PWBN projections for non-native ranges are, however, unrealistic biologically because a good portion of northern Canada, Greenland, Europe, and some parts of the Russian boreal forest are predicted to be suitable (Appendix S3, Fig. S7). Parthenium is from tropical and subtropical areas and therefore highly unlikely to be able to establish in boreal conditions, and our extensive search has not yielded a single record of the plant from these regions.

This strong mismatch between a very high AUC score and unrealistic projection maps indicated that there were severe problems with the traditional approach of computing AUC using withheld data from the training region (e.g., Peterson *et al*., 2007). When the model built from PWBN was evaluated under different

conditions (with both the background and presence points from across the world), the AUC scores dropped from 0.93 to 0.65, making PWBN the worst set of points for making models of global prediction. In fact, PWBN yielded the worst models for three of the five continents (Fig. 4a). The lack of background points from non-native ranges resulted in dramatic overprediction in non-native ranges (a situation that tends to increase AUC). Very few studies have quantitatively estimated model transferability (e.g., Mau-Crimmins *et al.*, 2006; Duncan *et al.*, 2009). We found that generalization and transferability of models (e.g., projecting invasive ranges outside of the training region) were best estimated quantitatively with AUC computed on distribution data from projected spaces (e.g., for each continent).

Previous studies have improved their models by including points from the invaded range and by partitioning the model prediction errors into various latitudinal bands in the western USA (Wenger & Olden, 2012). But to the best of our knowledge, no other study has taken our more complex approach of treating each continent as independent for the purposes of model building. We evaluated a model with occurrences and background points from each continent separately, and this approach provides a quantitative estimate of model transferability. This novel approach provides a unique method for improving projections into invaded ranges and thereby increasing model robustness.

Training regions. A small fraction of global grid cells have high habitat suitability for parthenium. From our ANOVA results, we observed a systematic decline in total variance with increasing suitability scores (Fig. 6a). When all grid cells were examined together, more variation in projected suitability was contributed by the point sources than by the SDM methods, with the relative importance of point source being even higher at habitats of low suitability. This indicates the importance of finding the best set of training points when making projections far from the current distribution of the invasive species. Conversely, all point sources tend to converge in their projection maps for the most highly suitable habitats (see Appendix S2-D for details).

We found that prediction accuracy was much improved using the global dataset for training the models (PWBW = presences from the world and background points from the world), rather than restricting training to the native range (PNBN = presences from native range and background points from native range), as also found by prior studies (Mau-Crimmins *et al.*, 2006; Broennimann & Guisan, 2008; Jiménez-Valverde *et al.*, 2011; Rodda *et al.*, 2011).

We showed that presences from different continents occupied different regions of environmental space (Fig. 3), as has been found in other studies of invasive species (Broennimann *et al.*, 2007; Beaumont *et al.*, 2009). Therefore, in order to encompass the set of environments that are suitable for parthenium, we needed to take presence points from the global distribution of the species. This result supports prior studies that have demonstrated that introduced ranges included in model training improve prediction in invaded ranges (Mau-Crimmins *et al.*, 2006; Broennimann & Guisan, 2008; Jiménez-Valverde *et al.*, 2011; Rodda *et al.*, 2011).

To understand why AUC computed in the traditional way (on the training region) performed poorly, we considered how AUC is computed. When PWBN models were tested on held-out data from the same ranges (presences from the world and background points *only* from native ranges), as the models attempted to maximize AUC scores, they ended up overpredicting outside the native ranges. But when these PWBN models were tested with background points from outside the native ranges, their AUC score decreased because most of the habitats considered suitable by the models were unsuitable in model testing data. Consequently, sensitivity (correctly predicting known occurrences) for PWBN stayed close to 1 outside the native ranges but specificity (correctly predicting the assumed absences) was between 0.05 and 0.19 (Fig. 7).

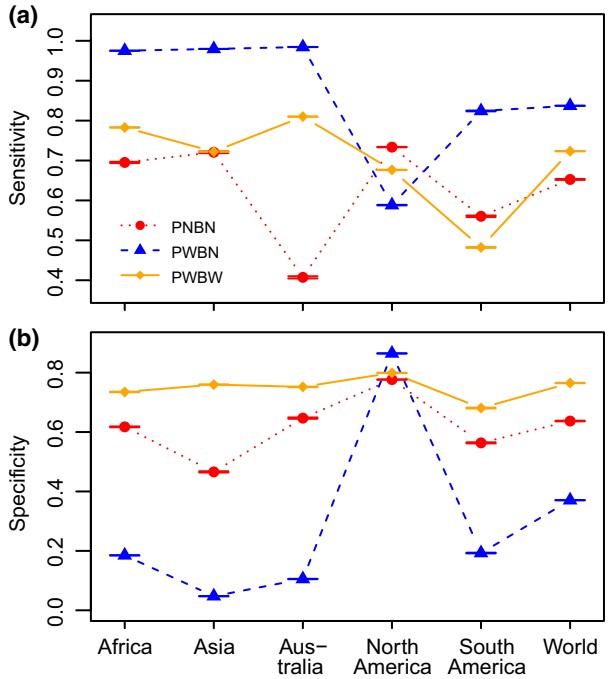

Fig. 7 Sensitivity (fraction of occurrence records predicted positive) and specificity (fraction of background points predicted negative) of the models built on three point sources. All models collapsed.

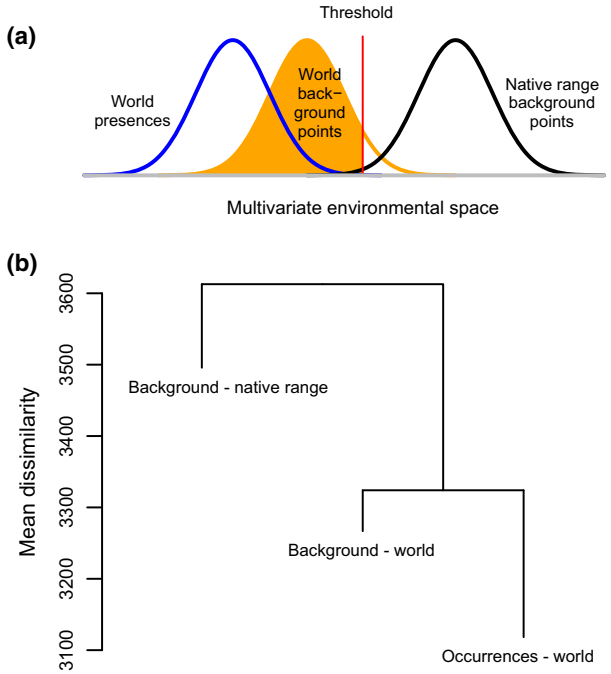

Fig. 8 (a) Illustration of our hypothesis that occurrence records from the world are more closely spaced in environmental space with background points from the world compared to background points from native range making models built with PWBN points highly inaccurate for prediction. Vertical red line represents the threshold in models built with presences from the world and background points from native range (PWBN). Whereas the threshold yields a very high AUC scores for PWBN models when evaluated with independent points from the training region, i.e., PWBN, it also classifies most of the environmental space of background points outside native ranges as positive inflating false-positive error rate (when evaluated with independent presences and background points from each continent) resulting in patterns of Fig. 7. (b) Pairwise Euclidean distances within and between groups showing how groups are spaced apart in multivariate environmental space (23 predictor variables; road excluded). The mean dissimilarity of 3324 between global presences and global background points is much smaller than the dissimilarity between global presences and native range background (3720). (The dendrogram shows the mean dissimilarity of native range background points with the other two groups together at slightly over 3600). Multiresponse permutation procedure (MRPP) shows that the groups differ significantly in the multivariate environmental space (P value < 0.001, A value = 0.0432, observed delta 3294, expected delta 3442).

To explain this result, we propose a hypothesis: In multivariate environmental space, global presences are more distant to native range background points than to world background points (Fig. 8a), allowing PWBN models to set a threshold that classifies the two types of points with the least amount of error (and therefore very high AUC) when tested with held-out data from training region. The hypothesized spacing of the clusters of points predicts that the broad environmental domain of presences in PWBN models includes most of the background points from invaded ranges; consequently, PWBN models, when evaluated with points from invaded ranges, yielded a very low specificity rate (0.05–0.19, Fig. 7). We computed Euclidean distances among the clusters of world presences, world background points, and native range background points in the environmental space of our 23 predictors (Fig. 8b). These distances supported our hypothesis that global presences are more environmentally similar to global background points than to native range background points. This explains why PWBN models, in spite of having the highest AUC scores in the model training space, have a very unrealistic prediction for non-native ranges (see Appendix S2-B for details).

Therefore, we dropped PWBN models from further consideration. Between PNBN and PWBW models, we chose PWBW for predicting the world; a very small gain in AUC (0.02) by PNBN models over PWBW models in native ranges is more than counterbalanced by a large gain in AUC (0.035–0.256) by PWBW models over PNBN models in non-native ranges.

Some recent studies have suggested that we may improve model reliability by focusing on efficient prediction of presences rather than absences (Phillips & Elith, 2010; Jiménez-Valverde *et al.*, 2011; Araújo & Peterson, 2012). However, we note that in the present study, this approach yielded unreliable models. Our PWBN models, with the highest AUC score on independent data from the training range (Fig. 4a, dashed box) and close to 100% accuracy in predicting presences (Fig. 7a), yielded very unrealistic projections at higher latitudes (Appendix S3, Fig. S7). This was most likely because the climatic niche of presences outside of the native range was not efficiently contrasted by the climatic space encompassing the pseudoabsences (discussed above).

Choosing SDM through combining information from standard metrics and biological insight. We observed that the projections of the four SDM methods outside the training region were substantially different, with some of them completely unrealistic (details in Appendix S2-C). Therefore, rather than build an ensemble projection (by averaging across the models), we chose the best projection(s) separately for each continent that best matched the biologically realistic expectations drawn from the expert opinion of our authors (details in Appendix S2-C). The model underlying that 'best' projection was the 'best' model.

There were some important differences in predictions made by the four SDM methods, the differences being

more dramatic further from the training region. For example, we concluded that GLM's predictions of highly suitable habitat in most of Greenland and part of northern Canada and northern Russia were very unrealistic (Appendix S3, Fig. S8). These regions have harsh winters that parthenium, a plant of tropical origin, cannot survive. Given that GLM was the only one of the four SDM methods to drastically deviate from expectations in areas that are far away from sampling regions, we believe that the extrapolation of GLM's parametric interaction terms between variables beyond the parameter space of model training is likely the cause. However, GLMs were commonly used in early analyses (Elith & Leathwick, 2009) and still are a widely used modeling method (Austin, 2002).

RF is a stronger classifier; compared to BRT, it has a tendency to overemphasize differences between grid cells. Its very flexible fitting procedure makes RF very effective in modeling complex responses (Berk, 2009). An unavoidable consequence of this flexibility that allows RF respond to highly local features of data is that it can inflate the risk of overfitting (Berk, 2009) and compromise its generalization, hampering its ability to make projections in a new landscape. BRT, on the other hand, reduces overfitting by giving different weight to the observations with highly local features, and averaging such fitting attempts. Essentially, this approach, called boosting, 'combines the outputs from many weak classifiers to produce a powerful committee' (Hastie *et al*., 2009). BRT, therefore, is likely to yield predictions that are more reliable outside of the training region than RF. These fundamental differences between RF and BRT match our observation: RF underpredicts Asia and southeastern Africa, and overpredicts South America and northern part of Africa including Sahara. Continent-wise, BRT gave the best predictions of all four modeling methods for Asia, North America, and South America; therefore, BRT not only secured one of the highest AUC scores but also closely matched our expectations about the species distribution. For Australia and Africa, GAM gave the best predictions (details in Appendix S2-C).

Expert opinion has been found to be useful in SDM (Murray *et al*., 2009). The importance of biological insight in model selection (details in Appendix S2-C) was heightened in the present study because the eight regions for which we computed AUC represented only 7.2% of the total grid cells on the planet for which projections were made.

Post hoc validation of our 'best' model in the field

Our *post hoc* test among SDMs used novel independent field data to validate projection outputs from models developed with entirely different datasets. One of the authors of this study (BBS) travelled extensively to collect distributional data of parthenium across Nepal in September and October 2013, after all of our models were completed. The 339 occurrence records he documented had a high correspondence with grid cells estimated to be suitable with our BRT model for Nepal. Our model projection was validated by the fact that the observed AUC of 0.76 (based on records collected after modeling) was statistically significantly different from the AUC expected under the null model (Fig. 9).

Future distribution of parthenium

In Asia, Africa, and South America, we identified vast stretches of highly suitable habitat for which no parthenium occurrences have been recorded. Eastern China, South-East Asia, and part of Japan and Korea were projected to harbor highly suitable habitat for the weed. In its native range, our projection maps suggested that parthenium was in equilibrium: Our results do not show large areas as suitable that are not already occupied. However, our results indicated that the archipelago that includes Cuba, Jamaica, Haiti, Dominican Republic, and Puerto Rico has high likelihood of being invaded by this weed as they provide highly suitable habitat, but our exhaustive search could obtain only seven occurrence records from that region.

Interestingly, our models showed that the coastal regions in the south (e.g., New South Wales) and west (e.g., Northern Territory) of Australia have some of the most suitable habitat for this weed. However, no major parthenium infestations are currently present in those areas. Even though we did not obtain a single occurrence record from that region, there have been cases of the weed being carried there by the flood events of 2010 and 2011. We believe this discrepancy between projected habitat suitability and lack of occurrence records is due to very effective management interventions to reduce, contain, or to eradicate parthenium where possible, in both states during the past several decades (Penna & MacFarlane, 2012). Also, strict quarantine measures are enforced across Australia for vehicle and grain movement from parthenium-infested areas. In addition, effective biological control and grazing management strategies have significantly reduced parthenium infestations in the core parthenium areas in central Queensland, resulting in reduced soil seed bank and limited the risk of parthenium seed spread to new areas (Dhileepan & McFadyen, 2012).

Africa, where several agencies are working toward the management of the weed, is likely to face stronger challenges. The entire eastern coastal belt of Africa, eastern half of Madagascar, Congo basin, coastal

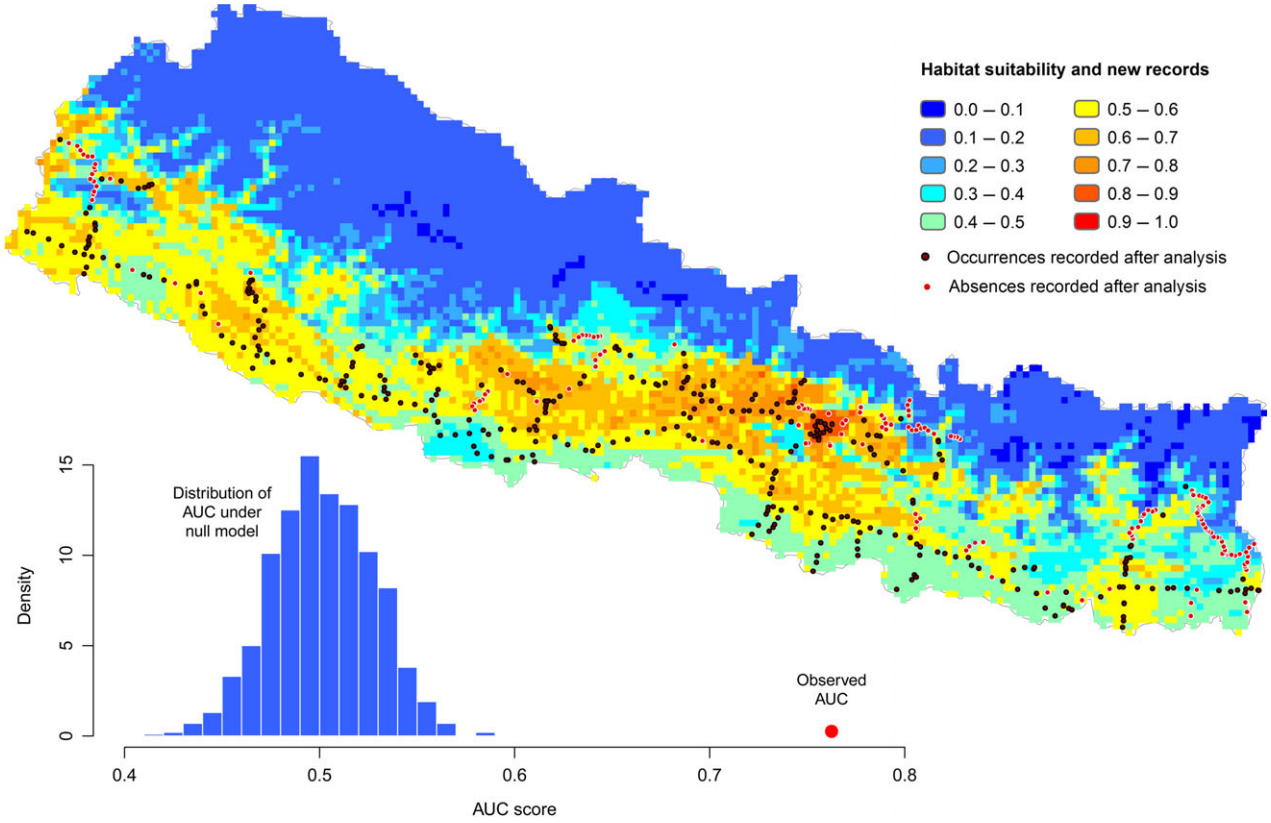

Fig. 9 *Post hoc* test of our best model. AUC was calculated with BRT prediction for grid cells in Nepal with 339 occurrence and 158 absence records, all collected after global projections were completed. This observed AUC (0.76) was statistically tested against AUC expected under a null model (Raes & ter Steege, 2007). The null AUC was computed 999 times with equal number of points (339 + 158 = 497) randomly drawn from the minimum area convex polygon encompassing the records, and randomly assigning the points to the category of 'present' or 'absent.' The observed AUC is significantly different from the distribution of AUC under null model (P value $\ll$ 0.001, t = −318.9291, one-tailed one-sample t-test).

regions of Ghana, and surrounding countries are projected to harbor highly suitable habitat for the spread and proliferation of parthenium.

Our habitat suitability projection roughly corresponds at a coarse spatial scale to the projection of parthenium with the use of CLIMEX model developed by McConnachie *et al.* (2011). However, our study differs substantially in both methodology and regional projections of suitability. McConnachie *et al.* constructed a single model from known climatic tolerances of parthenium and using its distribution in its native range and South Asia for making global projection models. In comparison, our approach used region-specific model selection and conducted continental cross-validation. Compared to our projection, McConnachie *et al.* (i) overpredicted the extent of suitable habitat in South America and Africa, (ii) underpredicted in eastern China, and (iii) projected the world at two orders of magnitude coarser spatial resolution, making it problematic to use their results for management interventions.

In summary, we found that construction of a highly reliable model for projecting future parthenium invasion potential required that (i) all geographic spaces were included in model training, (ii) flexible, data-defined smoothers were included to model nonlinear responses, and (iii) interactions between variables were modeled as they were discovered in data. We found that data-driven models, such as boosted regression trees, that (i) efficiently fit the dominant pattern but exclude highly local patterns in datasets and (ii) capture interactions as they appear in data rather than making *a priori* assumptions led to improved generalization of global projections of current distributions and hence improved projections of potential spread of parthenium.

Acknowledgements

Modeling was performed at Texas Advanced Computing Center (TACC). We thank John Fonner at TACC for his assistance with operating the supercomputer. We also thank agencies and people for their kind cooperation in supplying us with ~4000

occurrence records (Appendix S1, Table S1). Matthew Moskwik provided helpful comments on an early version of the manuscript; Abhishek Nakarmi downloaded occurrence records from some public domain sources; Iqbal Zeberi and Maan Rokaya provided their original occurrence records. This work was supported in part by National Science Foundation Earth Systems Modeling grant #1049208, the International Foundation of Science (IFS), Sweden (Grant no. C-5306) and ARC DECRA award # DE140101675.

Conflict of interest

None.

References

Acevedo P, Jiménez-Valverde A, Lobo JM, Real R (2012) Delimiting the geographical background in species distribution modelling. *Journal of Biogeography*, **39**, 1383–1390.

Ackerly DD (2003) Community assembly, niche conservatism, and adaptive evolution in changing environments. *International Journal of Plant Science*, **164**, S165–S184.

Adkins S, Shabbir A (2014) Biology, ecology and management of the invasive parthenium weed (*Parthenium hysterophorus* L.). *Pest Management Science*, **70**, 1023–1029.

Allouche O, Tsoar A, Kadmon R (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, **43**, 1223–1232.

Araújo MB, Guisan A (2006) Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, **33**, 1677–1688.

Araújo MB, Peterson AT (2012) Uses and misuses of bioclimatic envelope modeling. *Ecology*, **93**, 1527–1539.

Araújo MB, Whittaker RJ, Ladle RJ, Erhard M (2005) Reducing uncertainty in projections of extinction risk from climate change. *Global Ecology and Biogeography*, **14**, 529–538.

Auld BA, Hosking J, McFadyen RE (1982–83) Analysis of the spread of tiger pear and parthenium weed in Australia. *Australian Weeds*, **2**, 56–60.

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

Barbet-Massin M, Jiguet F, Albert CH, Thuiller W (2012) Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, **3**, 327–338.

Beaumont LJ, Gallagher RV, Thuiller W, Downey PO, Leishman MR, Hughes L (2009) Different climatic envelopes among invasive populations may lead to underestimations of current and future biological invasions. *Diversity and Distributions*, **15**, 409–420.

Berk RA (2009) *Statistical Learning from a Regression Perspective*. Springer-Verlag, New York, NY.

Boria RA, Olson LE, Goodman SM, Anderson RP (2014) Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, **275**, 73–77.

Box GEP, Draper NR (1987) *Empirical Model-Building and Response Surfaces*, p. 424, John Wiley & Sons, Inc., New York, NY.

Broennimann O, Guisan A (2008) Predicting current and future biological invasions: both native and invaded ranges matter. *Biology Letters*, **4**, 585–589.

Broennimann O, Treier UA, Müller-Schärer H, Thuiller W, Peterson AT, Guisan A (2007) Evidence of climatic niche shift during biological invasion. *Ecology Letters*, **10**, 701–709.

Cutler DR, Edwards TC Jr, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. *Ecology*, **88**, 2783–2792.

Dhileepan K (2009) Managing parthenium weed across landscapes: limitations and prospects. In: *Management of Invasive Weeds* (ed. Inderjit S), pp. 227–260, Springer Science, Dordrecht, Netherlands..

Dhileepan K (2012) Reproductive variation in naturally occurring populations of the weed *Parthenium hysterophorus* (Asteraceae) in Australia. *Weed Science*, **60**, 571–576.

Dhileepan K, McFadyen RE (2012) *Parthenium hysterophorus* L. – parthenium. In: *Biological Control of Weeds in Australia: 1960 to 2010* (eds Julien M, McFadyen RE, Cullen J), pp. 448–462. CSIRO Publishing, Melbourne, Vic.

Duncan RP, Cassey P, Blackburn TM (2009) Do climate envelope models transfer? A manipulative test using dung beetle introductions. *Proceedings of the Royal Society B: Biological Sciences*, **276**, 1449–1457.

Elith J, Graham C (2009) Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography*, **32**, 1–12.

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

Elith J, Graham CH, Anderson RP (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

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

Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ (2011) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, **17**, 43–57.

Fitzpatrick MC, Weltzin JF, Sanders NJ, Dunn RR (2007) The biogeography of prediction error: why does the introduced range of the fire ant over-predict its native range? *Global Ecology and Biogeography*, **16**, 24–33.

Graham GC, Lang CL (1998) *Genetic Analysis of Relationship of Parthenium Occurrences in Australia and Indications of its Origins*. Cooperative Research Centre for Tropical Pest Management, University of Queensland, Brisbane, Qld, Australia.

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

Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd edn. Springer-Verlag, New York, NY.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Jeschke JM, Strayer DL (2008) Usefulness of bioclimatic models for studying climate change and invasive species. *Annals of the New York Academy of Sciences*, **1134**, 1–24.

Jiménez-Valverde A, Peterson AT, Soberón J, Overton JM, Aragón P, Lobo JM (2011) Use of niche models in invasive species risk assessments. *Biological Invasions*, **13**, 2785–2797.

Kearney M, Phillips BL, Tracy CR, Christian KA, Betts G, Porter WP (2008) Modelling species distributions without using species distributions: the cane toad in Australia under current and future climates. *Ecography*, **31**, 423–434.

Kolar CS, Lodge DM (2001) Progress in invasion biology: predicting invaders. *Trends in Ecology & Evolution*, **16**, 199–204.

Lavergne S, Molofsky J (2007) Increased genetic variation and evolutionary potential drive the success of an invasive grass. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 3883–3888.

Lobo JM, Jiménez-Valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

Loo SE, Mac Nally R, Lake PS (2007) Forecasting New Zealand Mudsnail invasion range: model comparisons using native and invaded ranges. *Ecological Applications*, **17**, 181–189.

Mandle L, Warren DL, Hoffmann MH, Peterson AT, Schmitt J, von Wettberg EJ (2010) Conclusions about niche expansion in introduced *Impatiens walleriana* populations depend on method of analysis. *PLoS ONE*, **5**, e15297.

Mau-Crimmins TM, Schussman HR, Geiger EL (2006) Can the invaded range of a species be predicted sufficiently using only native-range data? Lehmann lovegrass (*Eragrostis lehmanniana*) in the southwestern United States. *Ecological Modelling*, **193**, 736–746.

McConnachie AJ, Strathie LW, Mersie W *et al.* (2011) Current and potential geographical distribution of the invasive plant *Parthenium hysterophorus* (Asteraceae) in eastern and southern Africa. *Weed Research*, **51**, 71–84.

Monahan WB (2009) A mechanistic niche model for measuring species' distributional responses to seasonal temperature gradients. *PLoS ONE*, **4**, e7921.

Murray JV, Goldizen AW, O'Leary RA, McAlpine CA, Possingham HP, Choy SL (2009) How useful is expert opinion for predicting the distribution of a species within and beyond the region of expertise? A case study using brush-tailed rock-wallabies Petrogale penicillata. *Journal of Applied Ecology*, **46**, 842–851.

Navie SC, McFadyen RE, Panetta FD, Adkins SW (1996) The biology of Australian weeds 27. *Parthenium hysterophorus* L. *Plant Protection Quarterly*, **11**, 76–88.

Navie SC, Panetta FD, McFadyen RE, Adkins SW (1998) Behaviour of buried and surface-lying seeds of parthenium weed (*Parthenium hysterophorus* L.). *Weed Research*, **38**, 338–341.

Parendes LA, Jones JA (2000) Role of light availability and dispersal in exotic plant invasion along roads and streams in the HJ Andrews Experimental Forest, Oregon. *Conservation Biology*, **14**, 64–75.

Pearman PB, Guisan A, Broennimann O, Randin CF (2008) Niche dynamics in space and time. *Trends in Ecology and Evolution*, **23**, 149–158.

Pearson RG, Dawson TP (2003) Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography*, **12**, 361–371.

Pearson RG, Thuiller W, Araujo MP *et al.* (2006) Model-based uncertainty in species range prediction. *Journal of Biogeography*, **33**, 1704–1711.

Penna A-M, MacFarlane M (2012) Parthenium incident in the Pilbara, Western Australia: how is this a good new story? In: *Proceedings of the 18th Australasian Weeds Conference* (ed. Eldershaw V), pp. 13–16. Weed Society of Victoria, Melbourne, Vic., Australia.

Peterson AT, Papeş M, Eaton M (2007) Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. *Ecography*, **30**, 550–560.

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

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Raes N, ter Steege H (2007) A null-model for significance testing of presence-only species distribution models. *Ecography*, **30**, 727–736.

Rao RS (1956) Parthenium – a new record for India. *Journal of Bombay Natural History Society*, **54**, 218–220.

Rodda GH, Jarnevich CS, Reed RN (2011) Challenges in identifying sites climatically matched to the native ranges of animal invaders. *PLoS ONE*, **6**, e14670.

Soberón J (2007) Grinnellian and Eltonian niches and geographic distributions of species. *Ecology Letters*, **10**, 1115–1123.

Soberón J, Peterson AT (2005) Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics*, **2**, 1–10.

Syfert MM, Smith MJ, Coomes DA (2013) The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. *PLoS ONE*, **8**, e55158.

Thuiller W (2003) BIOMOD – optimizing predictions of species distributions and predicting potential future shifts under global change. *Global Change Biology*, **9**, 1353–1362.

Thuiller W, Richardson DM, Pyšek P, Midgley GF, Hughes GO, Rouget M (2005) Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. *Global Change Biology*, **11**, 2234–2250.

Thuiller W, Albert C, Araújo MB, Berry PM, Cabeza M, Guisan A, Zimmermann NE (2008) Predicting global change impacts on plant species' distributions: future challenges. *Perspectives in Plant Ecology, Evolution and Systematics*, **9**, 137–152.

Thuiller W, Lafourcade B, Engler R, Araújo MB (2009) BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography*, **32**, 369–373.

Tyser RW, Worley CA (1992) Alien flora in grasslands adjacent to road and trail corridors in Glacier National Park, Montana (USA). *Conservation Biology*, **6**, 253–262.

Wenger SJ, Olden JD (2012) Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, **3**, 260–267.

Zurell D, Elith J, Schröder B (2012) Predicting to new environments: tools for visualizing model behaviour and impacts on mapped distributions. *Diversity and Distributions*, **18**, 628–634.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Additional tables (Tables S1–S5).
Table S1. Sources of occurrence points.
Table S2. Environmental predictors used in species distribution modeling.
Table S3. Sources of and explanation for using non-climatic variables.
Table S4. Four way analysis of variance for effect of various factors on AUC score.
Table S5. AUC scores for various regions computed with models using three point sources.
Appendix S2. Role of road and additional explanation.
Part A. Assessing the role of roads (including Figures S1–S3).
Fig. S1 Raster layers of distance to road.
Fig. S2 Effect of bias and road as explanatory variable on model performance (AUC $\pm$ 1 SE).
Fig. S3 Comparing the effect of including road in the set of predictors on projection: predictions of habitat suitability for India and surrounding by boosted regression trees (BRT).
Part B. Why AUC computed in traditional way ($AUC_{\text{training region}}$) performed poorly.
Part C. Which SDM method to select?
Part D. Sources of variation in projections.
Appendix S3. Additional figures (Figures S4–S8).
Fig. S4. Model evaluation with Cohen's kappa for the scenario PWBW, AreaBias, NoRoad.
Fig. S5. Replication of Fig. 5 in the paper (global projection of habitat suitability) with and without occurrence records for easy comparison.
Fig. S6. Relationship between various indices of model evaluation with pearson correlation coefficient displayed.
Fig. S7. Effect of point source: projection of habitat suitability for the world by boosted regression trees (BRT) using PWBN points.
Fig. S8. Unreliable projection of generalized linear models (GLM).