Weed Science, 51:863–868. 2003

Predicting the potential invasive distributions of four alien plant species in North America

A. Townsend Peterson
Corresponding author. Natural History Museum and Biodiversity Research Center, The University of Kansas, Lawrence, KS 66045; town@ku.edu

Monica Papes
Daniel A. Kluza
Natural History Museum and Biodiversity Research Center, The University of Kansas, Lawrence, KS 66045

Ecological niche modeling, a new methodology for predicting the geographic course of species' invasions, was tested based on four invasive plant species (garlic mustard, sericea lespedeza, Russian olive, and hydrilla) in North America. Models of ecological niches and geographic distributions on native distributional areas (Europe and Asia) were highly statistically significant. Projections for each species to North America—effectively predictions of invasive potential—were highly coincident with areas of known invasions. Hence, in each case, the geographic invasive potential was well summarized in a predictive sense; this methodology holds promise for development of control and eradication strategies and for risk assessment for species' invasions.

Nomenclature: Hydrilla, *Hydrilla verticillata* (L.f.) Royle HYLLI; Russian olive, *Elaeagnus angustifolia* L. ELGAN; sericea lespedeza, *Lespedeza cuneata* (Dum.-Cours.) G. Don LESCU; garlic mustard, *Alliaria petiolata* (Bieb) Cavara & Grande ALAPE.

Key words: Invasive species, ecological niche modeling, Genetic Algorithm for Rule-set Prediction, prediction.

Ecological niche modeling has seen considerable exploration as a tool for understanding complex biodiversity phenomena (Jones and Gladkov 1999; Joseph and Stockwell 2000; Peterson 2001; Peterson et al. 1999, 2000, 2001, 2002; Scott et al. 1996, 2002; Walker and Cocks 1991). On the basis of previous explorations (Zalba et al. 2000), Peterson and Vieglais (2001) presented the framework of a methodology for application of this tool to the challenge of predicting the geographic potential of species' invasions. This approach uses an ecological niche model based on the ecological characteristics of known occurrences in the native distribution of a species to identify suitable areas for the species on a potentially invaded range.

The application of ecological niche modeling in predicting species' invasions was introduced based on example tests with two avian invasions that occurred in North America in the past (Peterson and Vieglais 2001). The method has also been used to evaluate a series of additional invasions, including the Asian longhorned beetle (*Anoplophora glabripennis*) in North America (Peterson and Vieglais 2001), bass in Japan (Iguchi et al. 2003), and eastern North American owls in western North America (Peterson and Robins 2003). Nevertheless, the method has not been applied broadly to invasive plant species, particularly as regards economically important invasive pest species.

In this article, we develop retrospective tests for the predictive accuracy of ecological niche models for four invasive plant species in North America: garlic mustard, sericea lespedeza, Russian olive, and hydrilla. Each of the species is well established as an invasive species in North America. This suite of species thus offers ample opportunity for detailed statistical testing of the predictive abilities of the ecological niche–modeling approach.

Methods

Georeferenced occurrence points from the species' native ranges were drawn from diverse sources, including herbarium specimen records and scientific literature, including floras, systematic treatments, etc. Overall, 143 points were available for garlic mustard, 41 for Russian olive, 30 for hydrilla, and 28 for sericea lespedeza.

Ecological niches were modeled using the Genetic Algorithm for Rule-set Prediction (GARP) (Stockwell 1999; Stockwell and Noble 1992; Stockwell and Peters 1999). In general, the procedure focuses on modeling ecological niches (the conjunction of ecological conditions within which a species is able to maintain populations without immigration) (Grinnell 1917). Specifically, GARP relates ecological characteristics of known occurrence points to those of points sampled randomly from the rest of the study region, seeking to develop a series of decision rules that best summarize those factors associated with the species' presence (Feria and Peterson 2002).

Occurrence points are divided into training and test data sets—50% of the data points are set aside for a completely independent test of model quality (extrinsic test data), 25% are used for developing models (training data), and 25% are used for tests of model quality internal to GARP (intrinsic test data). Because these subsamples are made independently and randomly for each model run, GARP models effectively take advantage of most of the information in the input data set. GARP works in an iterative process of rule selection, evaluation, testing, and incorporation or rejection: a method is chosen from a set of possibilities (e.g., logistic regression, bioclimatic rules), applied to the training data, and a rule is developed or evolved. Predictive accuracy is then evaluated based on 1,250 points resampled from the test data and 1,250 points sampled randomly from the study region as a whole. Rules may evolve by a number of means that mimic DNA evolution: point mutations, deletions, crossing over, etc. The change in predictive accuracy from one iteration to the next is used to evaluate whether a particular rule should be incorporated into the model, and the algorithm runs either to 1,000 iterations or until convergence.

All modeling in this study was carried out on a desktop implementation of GARP now available for public download (Scachetti-Pereira 2001). This implementation offers much-improved flexibility in the choice of predictive environmental–ecological geographic information system (GIS) data layers: in this case, initially, we used 15 layers summarizing elevation, slope, aspect, flow accumulation (= upstream area contributing to water flow), flow direction (= modeled direction of water flow), topographic index (= tendency to pool water) (all from the U.S. Geological Survey Hydro-1K data set) (USGS 2001), and aspects of climate including mean annual diurnal temperature range, mean annual number of frost days, mean annual precipitation, mean annual solar radiation, mean annual maximum temperature, mean annual minimum annual temperature, mean annual temperature, mean annual water vapor pressure, and mean annual number of wet days (1961–1990; from the Intergovernmental Panel on Climate Change) (IPCC 2001). The area of analysis was Europe and western Asia for Russian olive and garlic mustard, eastern Asia for sericea lespedeza, and southern and eastern Asia for hydrilla.

To reduce environmental data layers to just those that provide the highest predictive accuracy, we used a jackknife manipulation. We ran multiple iterations of models, omitting each data layer, or suites of data layers, systematically. We then calculated correlations between inclusion of each data layer in the model (coded binarily) and omission error (percentage of extrinsic test presence data not predicted as present) to detect data layers that contribute negatively to model performance when evaluated based on independent test data. Correlations of the order of $r > 0.05$ were considered indicative of data layers that detract from model quality; such layers were removed from further analyses. It is important to note that jackknife manipulations were performed solely on native distributions of species and thus do not affect the independent nature of the invaded-range tests presented herein.

For production of final models, we submitted the occurrence data to GARP, which used 25% of the input data to generate models to be refined and evaluated using the remaining points. Unlike previous applications, which either used single models to predict species' distributions (Peterson 2001; Peterson et al. 2002) or summed multiple models to incorporate model-to-model variation (Peterson and Vieglais 2001), we used a new procedure (Anderson et al. 2003) for choosing best subsets of models. The procedure is based on the observations that (1) models vary in quality, (2) variation among models involves an inverse relationship between errors of omission (leaving out true distributional area) and errors of commission (including areas not actually inhabited), and (3) best models (as judged by experts blind to error statistics) are clustered in a region of minimum omission of independent test points and moderate area predicted (an axis related directly to commission error). The relative position of the cloud of points relative to the two error axes provides an assessment of the relative accuracy of each model. To choose the best subsets of models, we (1) generated 100 replicate models by repeated random resampling of training and test data sets, (2) eliminated all models that had omission errors on the basis of independent test points, (3) calculated the average area predicted present among these zero-omission models, and (4) identified models that were within 1% of the overall average area predicted. Model quality for native range predictions was tested using the extrinsic test data: chi-square tests were used to compare the observed success in predicting the distribution of test points on the basis of that expected under a random model (proportional area $\times$ number of extrinsic test points estimates the expected number correctly predicted if the prediction were to be random with respect to the distribution of the test points).

Projection of the rule-sets for these models onto maps of North America provided predictions of potential distributions. Predictive accuracy was tested by means of the following manipulations: (1) known occurrences were tallied from the PLANTS National Database (USDA 2002) as county records for Russian olive, garlic mustard, and sericea lespedeza, and from the nonindigenous aquatic species information resource as U.S. Environmental Protection Agency hydrologic unit codes (HUCs) for hydrilla (USGS 2002); (2) known occurrences (number of known county or HUC occurrences henceforth referred to as N_k) were transferred to Arc (ESRI 2001) shapefiles; (3) the observed predictive success was counted as the proportion of the N_k counties or HUCs predicted present by all best-subset models; (4) N_k counties or HUCs were chosen at random from the attributes table of the Arc shapefile, and the number predicted present by all the best-subsets models was counted; (5) the previous step was repeated 100 times to build a distribution of randomized predictive accuracies; and (6) the observed success was compared with the distribution of randomized results to obtain an approximate probability value for how unexpectedly good the best-subsets prediction was (i.e., one-tailed probability).

Results and Discussion

The jackknife manipulations identified data layers that detracted from the predictive abilities of the algorithm. For example, for Russian olive, diurnal temperature range, aspect, annual mean minimum temperature, mean annual temperature, and wet days were highly correlated with high omission error (all $r > 0.05$); the remaining 10 data layers were used in further analyses. For garlic mustard, elevation, flow accumulation, annual mean precipitation, solar radiation, and annual mean and maximum temperatures were eliminated. For sericea lespedeza, elevation and mean maximum annual temperature were eliminated. For hydrilla, no data layers were eliminated.

Best-subsets models for native ranges of each species (Figure 1) were highly statistically significant. The chi-square tests, based on the independent extrinsic test data sets, indicated predictive ability far better than random models (garlic mustard, all $P \leq 1.06 \times 10^{-9}$; Russian olive, all $P \leq 1.51 \times 10^{-4}$; sericea lespedeza, all $P \leq 4.50 \times 10^{-10}$; hydrilla, all $P \leq 2.04 \times 10^{-4}$). Hence, all best-subsets models were highly predictive on native distributions, and for that reason, we proceeded to explore their predictions for invaded distributional areas in North America.

Projecting the native range models for each species to North America, a variety of potential distributional extents were observed (Figure 1), ranging from relatively small (hydrilla and sericea lespedeza) to quite large (Russian olive and garlic mustard). In each case, however, the observed degree of coincidence between the projection of the native range

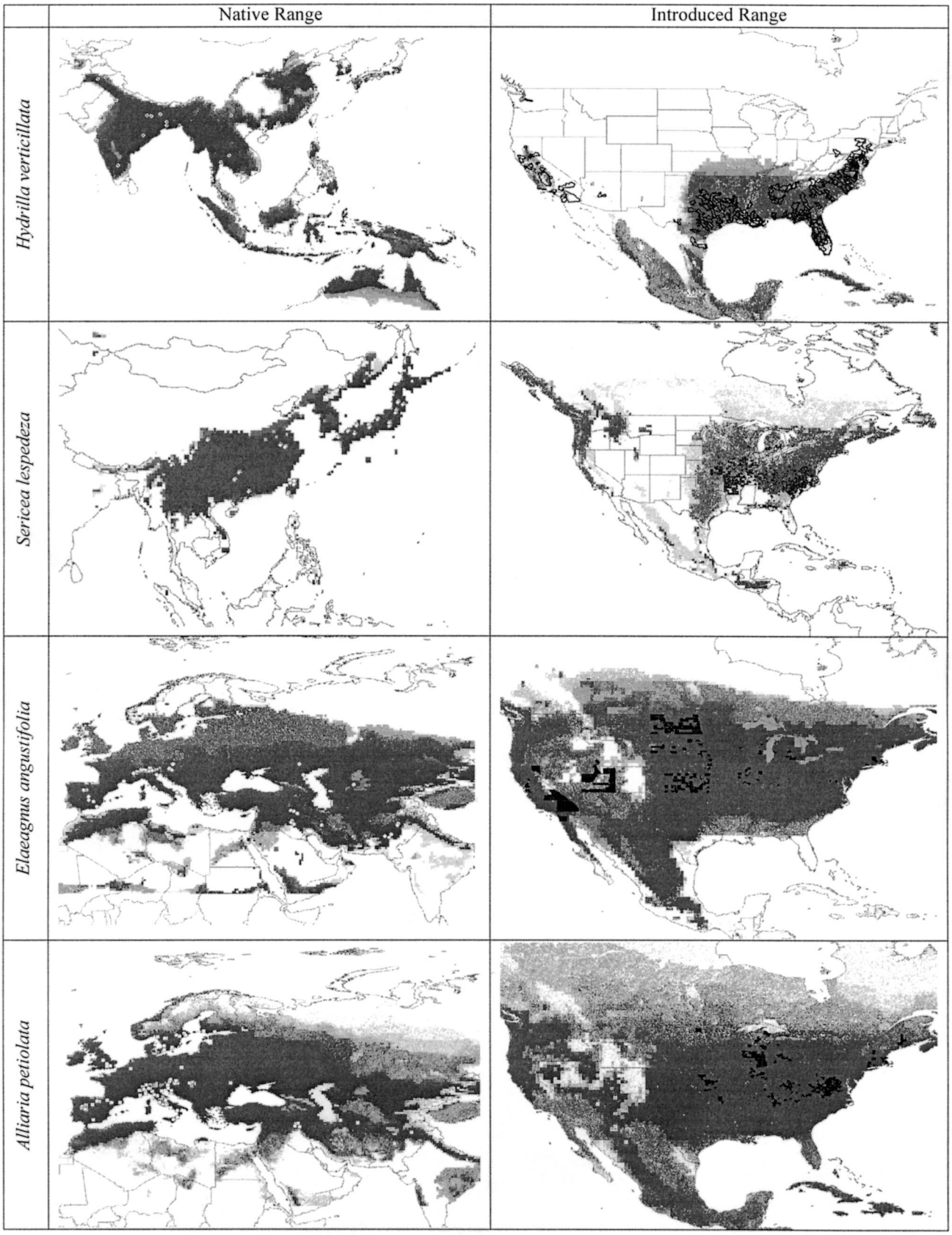

FIGURE 1. Predictions of native distributional areas and potential invaded distributional areas for hydrilla, sericea lespedeza, Russian olive, and garlic mustard. White symbols on native distribution maps indicate occurrence data used to build ecological niche models. Black polygons on the introduced distribution maps indicate known occurrences at the level of counties or hydrological units. Increasingly dark shading indicates greater confidence in prediction of presence (= model agreement).

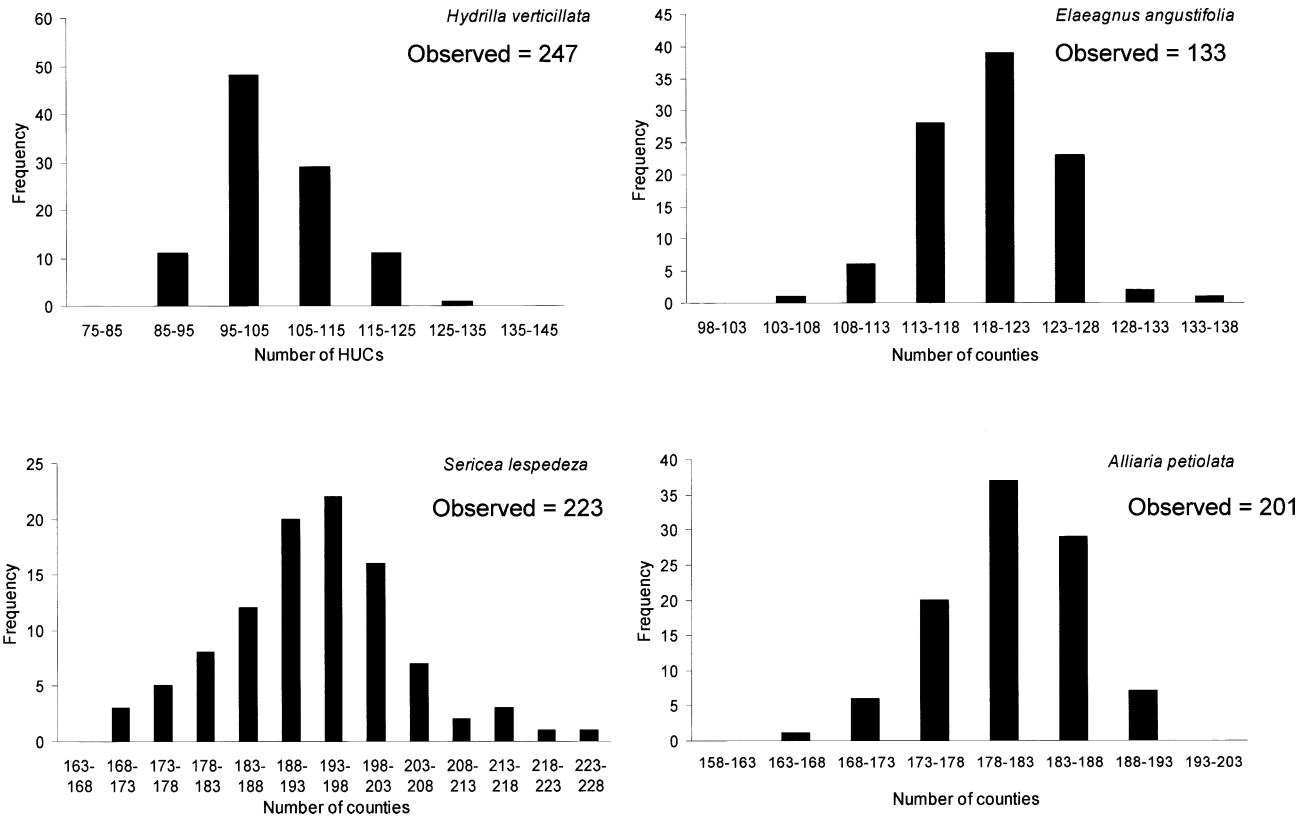

FIGURE 2. Results of validation tests for invaded distributional predictions for hydrilla, sericea lespedeza, Russian olive, and garlic mustard. Histograms indicate the frequency distributions of coincidence between model predictions and random suites of counties or hydrological units. Observed predictive success indicated for each species can be compared with these distributions to obtain an idea of model success in predicting occurrences above and beyond random expectations.

model and known occurrences was considerably better than the randomized replicates in the test of model prediction accuracy (Figure 2). As an example of tests of predictions for invaded ranges, for garlic mustard, random selections of 206 counties (the number of counties for which we have North American occurrence records) fell in areas of high prediction and produced a range of 161 to 190 counties; our model projections, however, successfully predicted 201 of the 206 known county occurrences, and hence the coincidence of our model predictions with actual invasion patterns was unexpectedly close ($P < 0.01$). For hydrilla also, the observed number of successful predictions (247) fell outside the range of randomized coincidences (90 to 130), indicating a significant prediction of the invaded range ($P \ll 0.01$). For Russian olive and sericea lespedeza, observed values fell in the highest category of the randomizations, well outside the 95th percentile of the distributions. Hence, in all cases, the ecological characteristics of species in their native geographic distributions successfully predicted the potential geographic extent of invasion in North America.

The results of this study are quite simple: ecological niche models developed on native geographic distributions and projected to other regions can predict the geographic potential of species' invasions with high accuracy. This conclusion echoes the original publications based on this approach or its parallels (Higgins et al. 1999; Holt and Boose 2002; Peterson and Vieglais 2001; Zalba et al. 2000). Beyond invasive applications, GARP's predictive abilities have been tested and proven under diverse circumstances (Anderson et al. 2002a, 2002b; Godown and Peterson 2000; Peterson 2001; Peterson and Cohoon 1999; Peterson and Vieglais 2001; Peterson et al. 1999, 2000, 2001, 2002; Stockwell and Peterson 2002).

The time required for this study, however, points to a critical bottleneck in the development of such predictive assessments. For any particular species, time was spent as follows: approximately 2 mo for accumulating native-distribution occurrence points; 1 h for invaded-distribution occurrence points; 3 d for jackknife manipulations; and 1 d for final model production, interpretation, and significance testing. In no case did we find a database that provided range-wide point information regarding occurrences in the native range.

To the extent that such predictive approaches are desirable, this study then points to a critical need for species' occurrence data. Herbarium data are relatively seldom in electronic form and are even more seldom available for download via the Internet. Indeed, in the *Species Analyst* (Vieglais 2000), a distributed biodiversity information network that serves more than 25 million biological specimen records, only data from the Canadian Royal Botanical Gardens, University of Kansas Herbarium, and the Arizona State University Herbarium are available for search and download. Additional herbaria are available for search on the Mexican national biodiversity server (CONABIO 2002); a few individual herbaria make their data available as well (e.g., California State University, Stanislaus; Museum of Evolution at Uppsala University, Sweden; Botanical Museum, University of Copenhagen, Denmark; and New York Botanical Garden). However, mainly, herbarium data are

quite difficult to access, a condition that will cause unending difficulties for the development of predictive models for invasive species and other applications.

Hence, access to point-occurrence information is complex at best. Moreover, once data are in hand, they still must be georeferenced, a process that can absorb significant amounts of time. A basic minimum requirement would be about 15 to 20 points well scattered throughout the species' native distribution—a previous analysis had indicated higher–sample size requirements (~ 50) (Stockwell and Peterson 2002), but recent advances in model development have made possible better model quality with fewer input occurrence points (Anderson et al. 2003). Development of models with even fewer input occurrence points is possible if one is willing to forego the step of model validation based on independent test occurrence data, thus using 50% of points for model development and 50% for model evolution.

A further challenge is the assembly of input environmental data sets. Although data comparable with those used herein are available from the sources cited above and are soon to be available as an environmental data packet for download with desktopGARP (Scachetti-Pereira 2001), numerous improvements are possible. For instance, multitemporal data from satellite sensors have shown excellent predictive ability in modeling species' native distributions (Egbert et al. 2002) and have shown promise in preliminary applications to species' invasions (A. T. Peterson et al., unpublished data). Nevertheless, remotely sensed data have yet to be applied and tested formally.

Data access considerations aside, however, this study further demonstrates the predictive power of ecological niche models for species' invasions. As has been indicated on the basis of independent lines of evidence (Peterson et al. 1999), species' ecological niches constitute long-term stable constraints on their distributional potential (Peterson and Vieglais 2001). This demonstration of predictivity of species' invasions further supports the conservatism hypothesis—species "obey" a consistent set of rules (the dimensions of the ecological niche) in their geographic distributions. This demonstration also lays the foundation for a new tool for investigators interested in anticipating and preventing successful species' invasions.

Ecological niche modeling thus offers a rich, new source of inferences and predictions regarding the geographic dimensions of species' invasions. Other algorithms have been applied to the challenge previously, such as climate envelopes (Holt and Boose 2002) or logistic regression (Higgins et al. 1999; Zalba et al. 2000). The GARP-based approach has the advantage of much-improved precision in its predictions: other approaches (particularly climate envelopes) tend to overpredict the dimensions of the niche rather drastically (Stockwell and Peterson 2002). This improved precision permits considerably improved predictive power for native distributions (Peterson 2001) and appears to translate directly into improved predictions of invasive potential.

Acknowledgments

Data were kindly provided by the Università "La Sapienza," Rome, Italy; Steiermärkisches Landesmuseum Joanneum, Graz, Austria; Departamento de Botánica, Universidad de Salamanca, Spain; and Moscow State University, Russia. Craig Freeman kindly provided distributional data from the Missouri Botanical Garden. Special thanks are due to Prof. G. G. Aymonin for his particularly helpful correspondence. This study was funded by the U.S. National Science Foundation and the U.S. Environmental Protection Agency.

Literature Cited

Anderson, R. P., M. Gomez, and A. T. Peterson. 2002a. Geographical distributions of spiny pocket mice in South America: insights from predictive models. Glob. Ecol. Biogeogr. Lett. 11:131–141.

Anderson, R. P., M. Laverde, and A. T. Peterson. 2002b. Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. Oikos 93:3–16.

Anderson, R. P., D. Lew, and A. T. Peterson. 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecol. Model. 162:211–232.

CONABIO. 2002. Red Mexicana de la Información de la Biodiversidad. www.conabio.gob.mx/.

Egbert, S. L., E. Martinez-Meyer, M. A. Ortega-Herta, and A. T. Peterson. 2002. Use of datasets derived from time-series AVHRR imagery as surrogates for land cover maps in predicting species' distributions. Pages 2337–2339 *in* Proceedings IEEE 2002 International Geoscience and Remote Sensing Symposium (IGARSS). Volume 4. Toronto, Canada.

ESRI. 2001. ArcView. Redlands, CA: Environmental Systems Research Institute.

Feria, T. P. and A. T. Peterson. 2002. Using point occurrence data and inferential algorithms to predict local communities of birds. Divers. Distrib. 8:49–56.

Godown, M. E. and A. T. Peterson. 2000. Preliminary distributional analysis of U.S. endangered bird species. Biodivers. Conserv. 9:1313–1322.

Grinnell, J. 1917. Field tests of theories concerning distributional control. Am. Nat. 51:115–128.

Higgins, S. I., D. M. Richardson, R. M. Cowling, and T. H. Trinder-Smith. 1999. Predicting the landscape-scale distribution of alien plants and their threat to plant diversity. Conserv. Biol. 13:303–313.

Holt, J. S. and A. B. Boose. 2002. Potential for spread of *Abutilon theophrasti* in California. Weed Sci. 48:43–52.

Iguchi, K., K. Matsuura, K. McNyset, A. T. Peterson, R. Scachetti-Pereira, D. A. Vieglais, E. O. Wiley, and T. Yodo. 2003. Predicting invasions of bass in Japan. J. Am. Fish. Soc. In press.

IPCC. 2001. Climate Data Archive. www.ipcc.ch/.

Jones, P. G. and A. Gladkov. 1999. FloraMap: A Computer Tool for Predicting the Distribution of Plants and Other Organisms in the Wild. Cali, Colombia: Centro Internacional de Agricultura Tropical. 99 p.

Joseph, L. and D.R.B. Stockwell. 2000. Temperature-based models of the migration of Swainson's flycatcher (*Myiarchus swainsoni*) across South America: a new use for museum specimens of migratory birds. Proc. Acad. Nat. Sci. Phila. 150:293–300.

Peterson, A. T. 2001. Predicting species' geographic distributions based on ecological niche modeling. Condor 103:599–605.

Peterson, A. T., L. G. Ball, and K. C. Cohoon. 2002. Predicting distributions of tropical birds. Ibis 144:e27–e32.

Peterson, A. T. and K. C. Cohoon. 1999. Sensitivity of distributional prediction algorithms to geographic data completeness. Ecol. Model. 117: 159–164.

Peterson, A. T., S. L. Egbert, V. Sánchez-Cordero, and K. P. Price. 2000. Geographic analysis of conservation priorities using distributional modelling and complementarity: endemic birds and mammals in Veracruz, Mexico. Biol. Conserv. 93:85–94.

Peterson, A. T. and C. R. Robins. 2003. When endangered meets invasive: ecological niche modeling predicts double trouble for spotted owls, *Strix occidentalis*. Conserv. Biol. In press.

Peterson, A. T., V. Sánchez-Cordero, J. Soberón, J. Bartley, R. H. Buddemeier, and A. G. Navarro-Siguenza. 2001. Effects of global climate change on geographic distributions of Mexican Cracidae. Ecol. Model. 144:21–30.

Peterson, A. T., J. Soberón, and V. Sánchez-Cordero. 1999. Conservatism of ecological niches in evolutionary time. Science 285:1265–1267.

Peterson, A. T. and D. A. Vieglais. 2001. Predicting species invasions using ecological niche modeling. Bioscience 51:363–371.

Scachetti-Pereira, R. 2001. Desktop GARP. www.lifemapper.org/desktopgarp.

Scott, J. M., P. J. Heglund, and M. L. Morrison, eds. 2002. Predicting

Species Occurrences: Issues of Accuracy and Scale. Washington, DC: Island. 840 p.
Scott, J. M., T. H. Tear, and F. W. Davis, eds. 1996. Gap Analysis: A Landscape Approach to Biodiversity Planning. Bethesda, MD: American Society for Photogrammetry and Remote Sensing. 320 p.
Stockwell, D.R.B. 1999. Genetic algorithms II. Pages 123–144 *in* A. H. Fielding, ed. Machine Learning Methods for Ecological Applications. Boston, MA: Kluwer.
Stockwell, D.R.B. and I. R. Noble. 1992. Induction of sets of rules from animal distribution data: a robust and informative method of analysis. Math. Comp. Simul. 33:385–390.
Stockwell, D.R.B. and D. P. Peters. 1999. The GARP modelling system: problems and solutions to automated spatial prediction. Int. J. Geogr. Informat. Syst. 13:143–158.
Stockwell, D.R.B. and A. T. Peterson. 2002. Effects of sample size on accuracy of species distribution models. Ecol. Model. 148:1–13.
USDA. 2002. Plants Database. www.plants.usda.gov/.
USGS. 2001. HYDRO1k Elevation Derivative Database. www.edcdaac.usgs.gov/gtopo30/hydro/.
USGS. 2002. Nonindigenous Aquatic Species. www.nas.er.usgs.gov/.
Vieglais, D. A. 2000. The Species Analyst. www.speciesanalyst.net/.
Walker, P. A. and K. D. Cocks. 1991. HABITAT: a procedure for modelling a disjoint environmental envelope for a plant or animal species. Glob. Ecol. Biogeogr. Lett. 1:108–118.
Zalba, S. M., M. I. Sonaglioni, and C. J. Belenguer. 2000. Using a habitat model to assess the risk of invasion by an exotic plant. Biol. Conserv. 93:203–208.

Received May 22, 2002, and approved April 03, 2003.