APPENDIX C. GARP MODEL VALIDATION

GARP MODEL VALIDATION

The most rigorous evaluation of any model is a test of its ability to correctly predict independent data that have not been processed by the model. In this report, Genetic Algorithm for Rule-set Production (GARP) models were evaluated by comparing the known occurrences of three invasive species already within the Great Lakes with the predictive ability of GARP models developed for these species using occurrence data from other regions.

Model performance was assessed using area under the curve of the receiver operating characteristic curve (Sing et al., 2005) using R 2.4.0 (R Development Core Team, 2006). Area under the curve is a threshold-independent evaluation of model performance that, in this case, measures the ability of the model to differentiate between sites where a species is considered present versus where it is considered absent. Area under the curve represents the probability that, when a predicted-present site and a predicted-absent site are drawn at random, the predicted-present site will have a higher predicted value than the predicted-absent site. Because true absence data were not available, randomly generated absence data, termed pseudo-absence data (i.e., points selected randomly from sites where the species have not been recorded as present within the Great Lakes), were used to validate the GARP models. This is a standard approach when true absence data are not available (Graham et al., 2004).

GARP produces predictions of habitat suitability ranging from 0 to 100 that can be converted to a binary prediction of presence or absence by selecting a threshold. For the purpose of model validation, values above this threshold (i.e., 50) are considered present (assigned a value of 1) while values below this threshold are considered absent (assigned a value of 0). The threshold that is selected is typically the threshold that maximizes model performance which may bias estimates of model performance. Area under the curve avoids the subjectivity in the threshold selection process and, therefore, provides an unbiased evaluation of model performance by plotting the false-positive rate (i.e., over-prediction, the rate at which the model predicts the species to be present at sites at which it is considered absent) versus the true-positive rate (i.e., the rate at which the model correctly predicts known presences as present) across *all* possible thresholds. For these reasons, area under the curve is considered one of the best approaches for model validation (Pearce and Ferrier, 2000). Nonetheless, area under the curve poses three important limitations. Notable to his study are that (1) it weights over-prediction and

C-2

under-prediction errors equally, (2) it does not give information about the spatial distribution of prediction errors, and (3) the size of the study area to which models are projected influences the rate of correctly predicted absences and the area under the curve scores.

Explanation of figures

The first figure, Figure C-1, is explained to help interpret the set of three figures. Figure C-1 shows the results from model validation for the zebra mussel, including a plot of the receiver operating characteristic curve with the area under the curve statistic. The colors along this curve correspond to the colors in the map of the zebra mussel predicted habitat suitability. Thus, by moving along the curve, one can stop at any color transition, say between yellow and orange (or a threshold value of 0.81 as determined by the right-hand y-axis or 81 from the legend in the map). By moving horizontally from this threshold value to the left-hand y-axis, one can determine the rate at which known presence correctly are predicted as present, also known as the true-positive rate (about 0.75 in this example). By moving vertically downward from this threshold to the x-axis, one can determine the rate at which pseudo-absences were predicted as present (the true-false rate, which is about 0.3 in this case). In other words, if values greater than 81 in the map are considered as present and values less than 81 as *absent*, we would get roughly 75% of the known occurrences correctly predicted, but it would also predict roughly 30% of presumed absences as present. In this manner, the receiver operating characteristic curve provides a means to assess the rates of false-positive predictions (predicting a species present where it is considered absent) and false-negative predictions (predicting a species absent where it is known to be present).

Model Evaluation Results

Swets (1988) suggested the following scale for determining model performance using area under the curve: 0.90-1.00 = excellent; 0.80-0.90 = good; 0.70-0.80 = fair; 0.60-0.70 = poor; $\leq 0.60 =$ fail. The area under the curve for all 3 species falls between 0.74 and 0.79, so the models would fall into the 0.70 to 0.80 category of "fair" (see Table C-1).

C-3

Table C-1. Summary of area under the curve (AUC) values and occurrence data sets used for model construction (training points) and evaluation points. Evaluation points within the Great Lakes are shown as hollow points on Figures C-1 to C-3

Species and common name	No. of Great Lakes evaluation points	No. and location of training points	AUC
Dreissena polymorpha	238	24 (Europe)	0.79
zebra mussel Gymnocephalus cernuus	46	183 (Europe)	0.79
rune <i>Potamopyrgus antipodarum</i> New Zealand mud snail	10	844 (Europe, Australia)	0.74

Model Evaluation

All three model validation Figures (C-1 to C-3) show (1) the predicted habitat suitability for each species within the Great Lakes when using only occurrence data from outside the Great Lakes, (2) the corresponding receiver operating characteristic curve plot and area under the curve value and bootstrap statistics for each model,

and (3) the occurrence data within the Great Lakes withheld from GARP and used for evaluation of predictive performance (hollow points). Note that for reasons discussed under "Selecting Species to Model and Development of Occurrence Data" in Section 3.1, these occurrence points may not be inclusive of all known occurrences in the Great Lakes and represent only those suitable for model evaluation.

Taken together, these area under the curve scores and the predicted distributions suggest three important conclusions. First, in our tests, GARP models adequately predict the known distributions of potential invasive species within the Great Lakes and, therefore, may be capable of accurately identifying areas of the Great Lakes susceptible to aquatic invasive species that have yet to be introduced. Thus, distribution data from a species' existing range can produce useful predictions of invasion potential using these GARP methods. Second, the observed patterns of invasion closely match those predicted for both *known* and *potential* invaders, suggesting that Lakes Erie and Ontario, near-shore areas of all of the Great Lakes in general, Saginaw Bay in Lake Huron, Lake St. Clair (located between Lakes Erie and Huron), and Thunder Bay in Lake Superior are particularly prone to future invasion when considering environmental tolerances alone. Finally, the universally high area under the curve scores suggest that the six environmental data layers we selected as inputs for the GARP models provide useful information for predicting the potential distributions of invasive species within the Great Lakes. In sum, the model validation exercise suggests that GARP predictions provide a useful assessment of invasion potential, given the availability of adequate occurrence data outside the Great Lakes.

REFERENCES FOR APPENDIX C.

Graham, CH; Ferrier, S; Huettman, F; et al (2004) New developments in museum-based informatics and applications in biodiversity analysis. TRENDS Ecol Evol 19(9): 497-503.

Pearce, J; Ferrier, S. (2000) Evaluating the predictive performance of habitat models developed using logistic regression. Ecol Model 133(3): 225–245.

R Development Core Team (2006). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available on-line at <u>http://www.R-project.org</u>.

Sing, T; Sander, O; Beerenwinkel N; et al. (2005) ROCR: visualizing classifier performance in R. Bioinformatics 21:3940–3941.

Swets, JA. (1988) Measuring the accuracy of diagnostic systems. Science 240:1285–1293.

Zebra mussel (Dreissena polymorpha)



Figure C-1. GARP model validation for zebra mussel showing predicted suitability and area under the curve (inset).



Figure C-2. GARP model validation for ruffe showing predicted suitability and area under the curve (inset).



Figure C-3. GARP model validation for New Zealand mud snail showing predicted suitability and area under the curve (inset).