APPENDIX B. Justification of modeling and variable selection approaches.

Separation

In attempting to fit occupancy models for our focal species within a maximum likelihood framework, we found evidence of considerable separation in models containing some of our candidate covariates. Separation is an inconvenient phenomenon apparent in discrete-response regression models that are fit by maximum likelihood, and that more or less perfectly predict the response outcomes (Albert and Anderson 1984, Hosmer and Lemeshow 2000). In the context of the state process of occupancy models, separation occurs when the linear predictor can be divided along a threshold, such that values all correspond to observed presences on one side of the threshold, to absences on the other. The likelihood function rises toward a discontinuity, and the values of model coefficients and standard errors are inflated, sometimes dramatically.

The irony of this problem is that in applying species distribution models to habitat gradients, perfect prediction is exactly the objective. We desired, therefore, to retain covariates for consideration even when they induced separation or quasi-separation. This effectively precluded our ability to fit models via maximum likelihood, since we could neither estimate parameters and standard errors, nor use likelihood-based model selection criteria. For this reason, we instead fit all occupancy models in the study within a Bayesian framework (Royle and Dorazio 2008), with estimation achieved via Markov Chain Monte Carlo sampling implemented in JAGS (Just Another Gibbs Sampler 3.3; Plummer 2012a) and called from R (R Core Team 2013) with the rjags package (Plummer 2012b). We used uninformative priors for all regression

coefficients in both ψ and *p* models. Convergence of parameters was checked with the Gelman-Rubin convergence diagnostic (Gelman and Rubin 1992) in the coda package (Plummer 2006).

Bayesian variable selection

Our initial set of predictors for each species included many redundant variables, a result of our stated modeling objectives. The oft-cited guidelines for information-theoretic model selection by Burnham and Anderson (2001, 2002) prescribe the use of a relatively small number of predictor variables, combined according to *a priori* hypotheses. However, we contend that many field surveys occupy a middle ground between "strong inference" and "data dredging", and that relegating such studies to the latter category is unrealistic and counterproductive. Nevertheless, authors who find themselves in this gray area may use intuition or expert opinion to reduce moderately large sets of variables -- all of which have some *a priori* support -- down to a manageable number (Steidl 2006, Sleep et al. 2007, Steidl 2007). Perhaps this lends such modeling exercises the appearance of adhering to Burnham and Anderson's (2001) guidelines. In our view, however, this step introduces considerable unacknowledged subjectivity, when a firm set of *a priori* hypotheses is legitimately lacking, since important variables or models may be jettisoned at the outset (Seoane et al. 2005). Correlated variables do not contain identical information, and we prefer a method of variable selection that is informed by data.

Given this preference for data-driven model selection, and our set of constraints and objectives, we chose to use a Bayesian variable selection technique (Kuo and Mallick 1998, Royle and Dorazio 2008:72).

LITERATURE CITED

Albert, A., J. Anderson (1984) On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71:1-10.

Burnham, K. P. and D. R. Anderson 2001. Kullback-Leibler information as a basis for strong inference in ecological studies. Wildlife Research. 28:111-119.

Burnham, K. P. and D. R. Anderson 2002. Model selection and multi-model inference: a practical information-theoretic approach. Springer, New York, New York, USA.

Gelman, A. and D. B. Rubin. 1992. Inference from iterative simulation using multiple sequences. Statistical Science. 7:457-472.

Hosmer Jr, D. W., S. Lemeshow, and R.X. Sturdivant. 2000. Applied Logistic Regression. Wiley, New York, New York, USA.

Kuo, L. and B. Mallick. 1998. Variable selection for regression models. Sankhyā: The Indian Journal of Statistics, Series B. 65:65-81.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. CODA: Convergence diagnosis and output analysis for MCMC. R News. 6:7-11.

Plummer, M. 2012a. JAGS Version 3.3.0 User Manual. http://mcmc-jags.sourceforge.net

Plummer, M. 2012b. rjags: Bayesian graphical models using MCMC. R package version 3-9. http://CRAN.R-project.org/package=rjags

Royle, J. A. and R. M. Dorazio. 2008. Hierarchical Modeling and Inference in Ecology. Academic Press, New York, New York, USA.

Seoane, J., J. Bustamante, and R. Díaz-Delgado. 2005. Effect of expert opinion on the predictive ability of environmental models of bird distribution. Conservation Biology. 19:512-522.

Sleep, D., M. Drever, and T. Nudds. 2007. Statistical versus biological hypothesis testing: response to Steidl. Journal of Wildlife Management. 71:2120-2121.

Steidl, R. J. 2006. Model selection, hypothesis testing, and risks of condemning analytical tools. Journal of Wildlife Management. 70:1497-1498.

Steidl, R. J. 2007. Limits of data analysis in scientific inference: reply to Sleep et al. Journal of Wildlife Management. 71:2122-2124.