Used-habitat calibration plots: A new procedure for validating species distribution, resource selection, and step-selection models

John R. Fieberg[1], James D. Forester[2], Garrett M. Street[3], Douglas H. Johnson[4], Althea A. ArchMiller[5], Jason Matthiopoulos[6]

[1]Department of Fisheries, Wildlife, and Conservation Biology, University of Minnesota-Twin Cities, 2003 Upper Buford Circle, Suite 135, St. Paul, MN 55108

[2]Department of Fisheries, Wildlife, and Conservation Biology, University of Minnesota-Twin Cities

[3]Department of Wildlife, Fisheries, and Aquaculture, Mississippi State University

[4]U.S. Geological Survey, Northern Prairie Wildlife Research Center

[5]Department of Fisheries, Wildlife, and Conservation Biology, University of Minnesota-Twin Cities

[6]Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow

Corresponding author: John R. Fieberg, Department of Fisheries, Wildlife, and Conservation Biology, University of Minnesota-Twin Cities, 2003 Upper Buford Circle, Suite 135, St. Paul, MN 55108. E-mail: jfieberg@umn.edu. Phone: 612-301-7132. Fax: 612-625-5299

Decision date: 05-Jun-2017

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: [10.1111/ecog.03123].

Abstract

"Species distribution modeling" was recently ranked as one of the top five "research fronts" in ecology and the environmental sciences by ISI's Essential Science Indicators (Renner and Warton 2013), reflecting the importance of predicting how species distributions will respond to anthropogenic change. Unfortunately, species distribution models (SDMs) often perform poorly when applied to novel environments. Compounding on this problem is the shortage of methods for evaluating SDMs (hence, we may be getting our predictions wrong and not even know it). Traditional methods for validating SDMs quantify a model's ability to classify locations as used or unused. Instead, we propose to focus on how well SDMs can predict the characteristics of used locations. This subtle shift in viewpoint leads to a more natural and informative evaluation and validation of models across the entire spectrum of SDMs. Through a series of examples, we show how simple graphical methods can help with three fundamental challenges of habitat modeling: identifying missing covariates, non-linearity, and multicollinearity. Identifying habitat characteristics that are not well-predicted by the model can provide insights into variables affecting the distribution of species, suggest appropriate model modifications, and ultimately improve the reliability and generality of conservation and management recommendations.

Keywords: animal movement, calibration, discrimination, inhomogeneous Poisson Process, logistic regression, prediction, presence-only, resource-selection, spatial point process, use-availability

Introduction

A variety of data collection and statistical methods are available for linking individuals, populations, and species to the habitats they occupy. Data collection methods range from design-based or opportunistic surveys that result in a set of pooled locations (ignoring any temporal component) (Edwards et al. 2006, Skov et al. 2016) to telemetry studies that result in many locations over time for a small number of individuals (Boyce and McDonald 1999, Pearce and Boyce 2006). A growing number of methods have been proposed for analyzing these different data types, and "species distribution modeling" (SDM) was recently ranked as one of the top five "research fronts" in ecology and the environmental sciences by ISI's Essential Science Indicators (Renner and Warton 2013). Regardless of the method used, the underlying objectives are the same: to understand how resources, risks, and environmental conditions influence distribution and abundance patterns (Mayor et al. 2009, Matthiopoulos et al. 2015). A more challenging, but equally important goal is to infer how various perturbations, including climate change and habitat management actions, influence these patterns (Matthiopoulos et al. 2011, Renner and Warton 2013). Unfortunately, SDMs frequently perform poorly when applied to novel environments (Elith et al. 2010, Matthiopoulos et al. 2011, Heikkinen et al. 2012, Wenger and Olden 2012). If we are going to use models to inform decision making, we need to have confidence in their predictions, which in turn requires that we have appropriate methods for model evaluation. Importantly, methods that provide insights into *why* a model performs poorly (e.g., missing predictors, incorrect

functional form, multicollinearity) are more useful than methods that provide only an overall measure of fit.

Much recent literature on model evaluation has focused on the interrelated concepts of model *validation*, *calibration*, and *discrimination* (Pearce and Ferrier 2000, Phillips and Elith 2010, Steyerberg et al. 2010, Harrell 2013, Chivers et al. 2014). Model *validation* is the process of assessing agreement between observations and fitted or predicted values. When a model (or set of models) is chosen via a data-driven process (e.g., transformations are considered, outliers are inspected and potentially dropped, and multiple models are compared before one or more are selected for inference), evaluations should ideally use out-of-sample data (i.e., data not used to arrive at the model(s); Araújo et al. 2005, Harrell 2013, Muscarella et al. 2014, Naimi and Araújo 2016). The use of out-of-sample data is also critical when evaluating model transferability and is especially challenging if the explanatory variables are correlated among themselves. Prediction error will typically be greater with the new data set unless the correlation among explanatory variables is the same as in the data originally used for model fitting (Dormann et al. 2013). When there is close agreement between observed and fitted/predicted values, we say the model is well *calibrated*; *calibration* therefore refers to steps taken to improve agreement between observed and predicted values (e.g., one may choose to 'shrink' regression parameters towards zero to improve out-of-sample predictions when models have been overfit; Harrell 2013, Street et al. 2016). *Discrimination*, by contrast, describes a model's ability to rank sample units in terms of their likely outcomes (Fielding and Bell 1997, Pearce and Ferrier 2000, Fawcett 2006, Steyerberg et al. 2010).

Calibration and discrimination often go hand-in-hand, though this need not be the case. A model may be well-calibrated but fail to discriminate well if it gives unbiased but highly imprecise estimates. A nice exemplification is given by Ellner et al. (2002), who demonstrated that estimates of extinction probabilities from population dynamic models are frequently too imprecise to rank individual populations in terms of risk even though they may provide an accurate estimate of the proportion of populations that will cross a quasi-extinction threshold. Conversely, a model may be poorly calibrated, yet have strong discriminating capabilities (Phillips and Elith 2010, Jiménez-Valverde et al. 2013). For instance, population indices may accurately rank sites in terms of their abundance, provided variation in detection probabilities is small relative to variation in abundance, even though indices are biased estimators of population size (Johnson 2008). Researchers routinely use methods such as the Area Under the Receiver Operating Curve (AUC) to evaluate discrimination of SDMs (e.g., Meyer and Thuiller 2006, Jiménez-Valverde 2012, Heikkinen et al. 2012), whereas calibration methods, the focus of this paper, are equally important but underutilized (Phillips and Elith 2010).

We consider methods for validating two general classes of models. The first includes a variety of methods appropriate for survey data pooled over time, in which observed locations are compared to a set of "background" (or "control" or "available") locations generated by randomly or systematically sampling from an area that encompasses the observed locations. Effectively, this approach treats the data as if they were cross-sectional (i.e., the temporal information in the data is ignored when making inferences). Animal telemetry data are also often analyzed in this way,

particularly when locations are collected infrequently or if the researcher is interested in habitat use at broad spatial scales (e.g., second or third orders of selection; Johnson 1980). Parallel development of methods for survey data and telemetry data has led to slightly different nomenclatures. The combination of the observed and random points is typically referred to as either presence-background (survey data) or use-availability (telemetry) data and the fitted models as either *species distribution models* (survey data) or *habitat-* or *resource-selection functions* or *models* (telemetry data). Though a variety of modeling approaches have been used in this context, most – MaxEnt (Elith et al. 2011), spatial logistic regression (Baddeley et al. 2010), weighted distribution theory with an exponential link function (Lele and Keim 2006), and resource utilization distributions (Millspaugh et al. 2006) – can be shown to be equivalent to fitting an inhomogeneous spatial point process model (Warton and Shepherd 2010, Aarts et al. 2012, Fithian and Hastie 2013, Hooten et al. 2013, Renner and Warton 2013).

The second class of models, developed for fine-scale telemetry data, also compares observed locations to a set of background points, but these background points are constrained to areas that are accessible to the animal from the previously observed location (a function of animal movement characteristics and sampling frequency). Each observed location is "paired" with a set of background/available points, resulting in highly stratified data. These data types are typically analyzed by fitting a conditional logistic regression (or equivalently, a discrete choice) model (Arthur et al. 1996, Manly et al. 2002), and the fitted models are referred to as *step-selection functions* (SSF) (Fortin et al. 2005, Forester et al. 2009, Thurfjell et al. 2014) or

integrated step-selection functions **(Avgar et al. 2016). Although these two classes of models share some features, calibration techniques developed for presence-absence (Harrell 2013) or presence-background data (Boyce et al. 2002, Johnson et al. 2006, Phillips and Elith 2010) do not easily generalize to step-selection functions because the data used to fit the latter models are highly stratified. Further, little work has been done to develop methods for validating step-selection models (but see Street et al. 2016).**

The popularity of SDMs, their propensity to fail when used to predict distributions in novel environments, and the current lack of sufficient diagnostics for evaluating models, especially those developed to analyze fine-scale telemetry data, are causes for concern. Here, we introduce a new method for model validation that can be applied across the entire spectrum of SDMs. Rather than focus on validating a binary response variable ($Y = 1$ for presence locations and 0 for background locations), we proposed to validate models by comparing distributions of the explanatory variables at the observed and predicted presence locations – i.e., the habitat characteristics associated with the used locations. These plots, which we refer to as *Used-Habitat Calibration plots* or UHC plots, complement existing approaches for validating traditional (non-stratified) species distribution or habitat selection models and also fill a void by providing a way to validate step-selection functions. Through a series of simulated and empirical examples, we show how UHC plots can help with three fundamental challenges of habitat modeling: identifying missing covariates, non-linearity, and multicollinearity.

Pooled-Survey Data Examples

We begin by considering two simple simulation examples where the variables influencing species distribution patterns are known. These examples are useful for testing if model validation tools return sensible and informative results under known model misspecifications. In particular, we will use these examples to explore the ability of model validation tools to diagnose a missing predictor or the need for a non-linear term. To understand the data-generating process, let $f^a(x)$ describe the available or background distribution of covariate(s) x in environmental space (i.e., $f^a(x)$ gives the relative frequency with which different values or levels of x occur across the entire landscape). Further, let $f^u(x)$ describe the distribution of the covariate(s) at used (i.e., presence) locations.

In our first example, constructed to explore the impact of a missing predictor, the species distribution was driven by elevation (x_1) and precipitation (x_2), with the species preferring sites at higher elevations and with lower levels of precipitation. In this example, the distribution of x_1 and x_2 in environmental space was assumed to be normal and centered to have mean 0: $f^a(x_1, x_2) = \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$. We considered three different data-generating scenarios in which we set var(x_1) = var(x_2) = 4, but varied cor(x_1, x_2) = ρ_{x_1,x_2} to explore how the effect of a missing predictor depends on the correlation among predictor variables. In the first scenario, we set $\rho_{x_1,x_2} = 0$ in both training and test data sets. In the second scenario, we set $\rho_{x_1,x_2} = -0.3$ in both training and test data sets, and in the third scenario, we set $\rho_{x_1,x_2} = 0.3$ in the training data set and $\rho_{x_1,x_2} = -0.3$ in the test data set. For each of these three scenarios, we

formed training data by choosing 100 presence locations, with the probability of selection proportional to $\exp(0.5x_1 - x_2)$. We combined these locations with a set of 10,000 randomly generated background points from $f^a(x_1, x_2)$. We set $Y = 1$ for the 100 presence locations and $Y = 0$ for the 10,000 background locations. We used the same approach to form a test data set of the same size (100 presence and 10,000 background locations).

We fit two different logistic regression models to the training data. First, we fit a model that included only elevation. Second, we fit a model that included both elevation and precipitation (the correct model). The estimated regression coefficients for elevation and precipitation were close to the data-generating values of 0.5 and -1 whenever we fit the correct model (i.e., $y \sim$ elev + precip; Table 1). The coefficient for elevation was also close to the data-generating value of 0.5 if we fit the model without precipitation, provided $\rho_{x_1,x_2} = 0$. By contrast, the coefficient for elevation in the model without precipitation was too high when $\rho_{x_1,x_2} = -0.3$ and too low when $\rho_{x_1,x_2} = 0.3$ (Table 1). This type of bias, referred to as *omitted-variable bias*, is well-known and is a function of $\text{cor}(x_1, x_2)$ and $\text{cor}(y, x_2|x_1)$ (Clarke 2005).

We considered a second example to explore the effect of model misspecification, where the species distribution exhibits a non-linear response to temperature (x_3). The optimal temperature for this species was set at $x_3 = 1$, with habitat suitability dropping off for warmer and colder temperatures. We again considered centered values of x_3, assumed to be normally distributed on the landscape with $f^a(x_3) = N(0, 4)$. We formed test and training data using the same approach as in the previous

example, but with the probability of selecting locations proportional to $\exp(2x_3 - x_3^2)$.

We fit a model with only a linear effect of temperature on the logit scale and another that also included a quadratic term (the correct model). The coefficient for temperature was too low when we fit the model with only temperature, but the coefficients were close to the data-generating values of 2 and −1 when both temperature and temperature^2 were included in the model (Table 2).

In subsequent sections, we evaluate each model's ability to predict presence locations in the test data. R code (R Core Team 2015) for generating the data and performing all analyses in the paper, along with any associated output, have been archived within the Data Repository for the University of Minnesota (accessible here: http://doi.org/10.13020/D6T590; Fieberg et al. 2016). We have also included functions for simulating and analyzing these data in an R package named *uhcplots* hosted on GitHub (Fieberg and ArchMiller 2016). This package can be downloaded using the install_github() function in the devtools library: devtools::install_github("aaarchmiller/uhcplots").

Calibration Plots

Methods for validating models include goodness-of-fit tests, diagnostic plots to assess model assumptions (e.g., residual versus fitted plots), and calibration plots of observed versus predicted values, where the latter are formed using cross-validation or bootstrapping (Phillips and Elith 2010, Harrell 2013). Calibration plots are particularly useful since they provide an honest measure of model fit by using

different data sets to fit and then evaluate the model. Unfortunately, calibration plots have received relatively little attention in the species distribution literature (but see Phillips and Elith 2010). Because many ecologists are unfamiliar with calibration plots, we will work towards our suggested approach by first detailing the steps necessary for producing a calibration plot when logistic regression is used to model binary (presence-absence) data. We then describe how calibration plots have been modified to work with presence-background data and illustrate these methods in conjunction with the above simulated data examples. With this foundation in place, we develop an alternative method of model calibration that focuses on the distribution of habitat characteristics at locations where the species is present.

Calibration Plot for Presence-Absence Data

Let Y represent the presence or absence of a species, a Bernoulli random variable with mean that is dependent on covariates X, $E[Y|X] = P(Y = 1|X) = \pi$. Further, let (x^{train}, y^{train}) refer to predictor and response data, respectively, used to fit the model and (x^{test}, y^{test}) refer to predictor and response data used to validate model predictions. In real applications, test and training data may be formed by data splitting, using k-fold cross-validation (Muscarella et al. 2014), or by sampling data with replacement multiple times (i.e., separate bootstrap samples; Harrell 2013, Fieberg and Johnson 2015). Alternatively, the model may be validated with data collected at another point in time or space, leading to a more stringent test of a model's predictive ability. To produce a calibration plot with presence-absence data:

1. **Estimate regression parameters, $\hat{\boldsymbol{\beta}}^{train}$, by fitting a logistic regression model to the *training* data $(\boldsymbol{x}^{train}, \boldsymbol{y}^{train})$.**
2. **Form predictions for the *test* data using $\boldsymbol{x}^{test}$ and the parameters estimated from the *training* data (i.e., $\hat{\boldsymbol{\beta}}^{train}$ from step [1]): $\hat{\boldsymbol{\pi}}^{test} = \frac{\exp(x^{test}\hat{\beta}^{train})}{1+\exp(x^{test}\hat{\beta}^{train})}$.**
3. **Form a calibration plot using one of three options:**
   - **Option 1: Bin the $\boldsymbol{y}^{test}$ data (e.g., based on quantiles of $\hat{\boldsymbol{\pi}}^{test}$). Plot the proportion of values where $\boldsymbol{y}^{test} = 1$ in each bin versus mean $\hat{\boldsymbol{\pi}}^{test}$ in each bin.**
   - **Option 2: Fit a new logistic regression model to the test data, considering a single predictor, $\boldsymbol{x}^{test}\hat{\boldsymbol{\beta}}^{train}$ (i.e., the logit of the predicted values): $\text{logit}(\boldsymbol{E}[\boldsymbol{Y}^{test}|\boldsymbol{X}^{test}]) = \boldsymbol{b_0} + \boldsymbol{b_1}(\boldsymbol{x}^{test}\hat{\boldsymbol{\beta}}^{train})$. Plot the fitted line with confidence intervals.**
   - **Option 3: Fit a more flexible, non-linear model (e.g., using regression or smoothing splines): $\text{logit}(\boldsymbol{E}[\boldsymbol{Y}^{test}|\boldsymbol{X}^{test}]) = \boldsymbol{f}(\boldsymbol{x}^{test}\hat{\boldsymbol{\beta}}^{train})$, and plot the fit of the model with confidence intervals.**

If the model is well-calibrated, we should see the binned values (option 1) or the fitted curves (options 2 and 3) line up well with the 1:1 line. Further, estimates of $(\boldsymbol{b_0}, \boldsymbol{b_1})$ should be close to (0, 1) (option 2) if the model is well-calibrated. If estimates of $(\boldsymbol{b_0}, \boldsymbol{b_1})$ are far from (0, 1), then one may choose to use $(\boldsymbol{b_0}, \boldsymbol{b_1})$ to re-calibrate the model (Giudice et al. 2012, Harrell 2013).

Calibration Plots for Presence-Background Data

Presence-background data differ from presence-absence data in that the zeros (the background data) may be utilized by the species (i.e., they are not 'true absences'). Boyce et al. (2002) and Johnson et al. (2006) developed a calibration plot for presence-background data that has been widely used to validate habitat selection models fit to telemetry data using logistic regression. Rather than use predicted probabilities from the fitted logistic regression model in step [2], Boyce et al. (2002) suggested using $w(x^{test}\hat{\beta}^{train}) = \exp(x^{test}\hat{\beta}^{train})$ for model calibration. Although this approach might at first appear to be ad hoc, it can be justified by recognizing that most methods for analyzing presence-background data, including logistic regression, can be shown to be equivalent to fitting an inhomogeneous Poisson process (IPP) model (Warton and Shepherd 2010, Aarts et al. 2012, Fithian and Hastie 2013, Hooten et al. 2013, Renner and Warton 2013). The likelihood for an IPP model, conditional on n_u total used (i.e., presence) locations from area A, is given by:

$$(1) \quad L(y_i|x_i,\beta) = \prod_{i=1}^{n_u} \frac{\exp(x_i\beta)}{\int_A \exp(x(s)\beta)ds}$$

The n_a randomly (or systematically) sampled available (i.e., background) points serve to approximate the integral in the denominator:

$$(2) \quad L(y_i|x_i,\beta) \approx \prod_{i=1}^{n_u} \frac{\exp(x_i\beta)}{\sum_{j=1}^{n_a+n_u} w_j \exp(x_j\beta)},$$

where the w_j are quadrature weights used to approximate the integral in eq. (1) using numerical integration techniques (ideally, the number of background points should be large enough that regression parameter estimators do not change with the addition of more points; Warton and Shepherd 2010). Thus, conditional on the set of used and

available points (n_u, n_a), the probability of selecting each point is proportional to $\exp(x\beta)$.

Boyce et al. (2002) and Johnson et al. (2006) suggested using k-fold cross-validation to form a binned calibration plot. After forming predictions via cross-validation, the plot is constructed via the following steps:

1. Bin the y^{test} data using quantiles of $w(x^{test}\hat{\beta}^{train})$ and calculate the mean value of $w(x^{test}\hat{\beta}^{train})$ in each bin, $\overline{w}_i$ (i = 1, 2 ..., n_{bins}).
2. Determine the number of used locations in each bin, n_u^i.
3. Determine the expected number of used locations in each bin, $E[n_u^i] = n_u^{test}\frac{\overline{w}_i}{\sum_{k=1}^{n_{bins}}\overline{w}_i}$, where n_u^{test} is the total number of used (i.e., presence) locations in the test data set. (Note: this equation can be modified slightly if the number of locations in each bin is not constant, see Johnson et al. 2006).
4. Plot n_u^i versus $E[n_u^i]$ along with a 1:1 line. As with presence-absence calibration plots, models with adequate fit should result in points that largely follow the 1:1 line.

Boyce et al. (2002) also advocated for calculating the Spearman correlation between n_u^i and $E[n_u^i]$. As noted by Phillips and Elith (2010), the Spearman correlation provides an alternative, non-parametric method for assessing calibration. Johnson et al. (2006) also suggested fitting a linear regression model relating n_u^i to $E[n_u^i]$, which should result in intercept and slope estimates close to 0 and 1, respectively, if the model is well-calibrated. Lastly, we note that Phillips and Elith (2010) proposed a

similar presence-background calibration plot using statistical smoothers to evaluate fit, thus avoiding the need to bin the data.

Application of Presence-Background Calibration Plots to Pooled-Survey Data Examples

Following Johnson et al. (2006), we constructed presence-background calibration plots for the models fit to each of the simulated pooled-survey data sets (Fig. 1, Fig. 2). In the first example, both models resulted in calibration plots that roughly followed the 1:1 line as long as ρ_{x_1,x_2} was the same in the test and training data (Fig. 1A−D). When ρ_{x_1,x_2} differed between the test and training data, the calibration plot for the elevation-only model differed significantly from the 1:1 line (Fig. 1E), whereas the correct model remained well-calibrated (Fig. 1F). Another noteworthy feature of the calibration plots, particularly those for the correct model (Fig. 1B, D, F) or the elevation-only model in the case where ρ_{x_1,x_2} = -0.3 for training and test data (Fig 1C), is a clustering of observed and expected counts near 0, except for the largest bin. This tight clustering reflects the high discriminatory ability of the models (i.e., they are able to clearly identify those points that have the highest relative probability of use). In the second example, the model containing only a linear effect of temperature resulted in a calibration plot with points that were widely scattered, and although the regression line was close to the 1:1 line, the R^2 is 0.04, suggesting the model did a poor job of predicting presence points in the test data (Fig. 2A). By contrast, the points in the calibration plot for the correct model, containing both temperature and temperature^2, closely followed the 1:1 line (R^2 = 0.99; Fig. 2B) suggesting this model was well-calibrated.

In summary, using presence-background calibration plots, we were able to correctly identify poorly calibrated models when we were missing an important predictor (but only when the correlation among predictor variables changed between training and test data sets; Fig. 1E) or when we needed to include a non-linear term (Fig. 2A). By themselves, however, these plots provide little additional insight into what might be causing the lack-of-fit or ways that the model might be improved.

Used-Habitat Calibration (UHC) Plot

A variety of residual plots (e.g., partial residual plots, added variable plots) have been developed to evaluate the potential for missing predictors or the need for non-linear terms in linear and generalized linear models (e.g., Kutner et al. 2005, Moya-Laraño and Corcobado 2008). Here, we develop a simple method for producing calibration plots that accomplish these same goals, but we use out-of-sample predictions. Specifically, we develop calibration plots that evaluate how well a model predicts the *characteristics* associated with the used (presence) locations. We call this type of plot a *Used-Habitat Calibration* plot (or UHC plot) and describe the steps for producing such plots below (see Fig. 3 for an illustration of the steps in the context of the first simulation example using the model with elevation but without precipitation).

Let x represent the full suite of explanatory variables included in the fitted model, n_u^{test} the total number of used (i.e., presence) locations in the test data set, and z the covariates of interest (these may be covariates already included in the model or additional covariates that may be under consideration for inclusion in the model).

The dimension of z may be greater than that of x, for example, if one chooses to begin with a simple model before progressively considering more complex models with additional covariates. Further, z may contain covariates that are available in the test data but are absent from the training data (e.g., if the model is applied to a new site where additional covariate data have been collected). In the example illustrated in Fig. 3, x includes only elevation, but z includes both elevation and precipitation.

1. **Summarize the distribution of z at the used (i.e., presence) points in the test data set, $f^u(z)$. In our examples, we use a kernel density estimator to represent $f^u(z)$ (solid black lines/density plots in Fig. 3; Wand and Jones 1994). Similarly, summarize the distribution of z at the available (i.e., background) points in the test data set, $f^a(z)$ (dashed red lines/density plots in Fig. 3). Differences between these two densities signal that the covariate will be an important predictor of the species distribution.**

2. **Fit a model to the training data set. Store $\hat{\boldsymbol{\beta}}$ and côv $(\hat{\boldsymbol{\beta}})$ to characterize the uncertainty in the parameters (ignoring the intercept if using logistic regression). Assuming we have a large enough sample for $\hat{\boldsymbol{\beta}}$ to be approximately normally distributed, we can draw samples from a multivariate normal distribution, $\mathbf{N}(\hat{\boldsymbol{\beta}}, \text{côv}\,(\hat{\boldsymbol{\beta}}))$, to account for uncertainty in the estimated parameters. This uncertainty may alternatively be captured using a non-parametric bootstrap or via samples from a posterior distribution (if implementing the model in a Bayesian framework); bootstrapping could also be used to account for parameter uncertainty in machine learning applications (e.g., models fit using random**

forests, artificial neural networks, etc.). We will refer to the distribution capturing uncertainty in $\hat{\boldsymbol{\beta}}$ as the *joint parameter distribution* to recognize that this will be a multivariate distribution if more than one covariate is included in the model.

3. Do the following M times (with loop index i):
   a. To account for parameter uncertainty, select new vector of parameter values randomly from their joint parameter distribution, $\boldsymbol{\beta}^i$.
   b. Estimate the relative probability of selection for the test data (given by eq. (2)): $w(\boldsymbol{x}^{test}\boldsymbol{\beta}^i) = \exp(\boldsymbol{x}^{test}\boldsymbol{\beta}^i)$.
   c. Select a simple random sample of n_u^{test} observations from the combined (presence and background) test data, with probabilities of selection proportional to $w(\boldsymbol{x}^{test}\boldsymbol{\beta}^i)$ from step [3b].
   d. Summarize the distribution of z associated with the points chosen in step [3c], $\hat{f}^u(z)_i$ (gray lines/density curves in Fig. 3).
4. Compare the observed distribution of covariate values at the presence points, $f^u(z)$ (black solid lines) from step [1], to the predicted distribution of these characteristics, $\hat{f}^u(z)_i$ (gray bands) from step [3], across the M simulations. One option is to overlay $f^u(z)$ (from step [1]) on a 95% simulation envelope constructed using the $\hat{f}^u(z)_i$ (Fig. 3). Alternatively, one might choose to plot the 2.5^{th} and 97.5^{th} quantiles of $f^u(z) - \hat{f}^u(z)_i$. We include functions in the *uhcplots* package for constructing these plots and illustrate the latter type of plot

in supplementary files archived with the Data Repository for the University of Minnesota (Fieberg et al. 2016, Fieberg and ArchMiller 2016).

Application of UHC Plots to Pooled-Survey Data Examples

To create UHC plots for the pooled-survey data examples, we constructed 1,000 predicted distributions of habitat covariates at the presence points in the test data set (i.e., $M = 1,000$ in step [3]) using the models fit to the training data, accounting for uncertainty in $\hat{\beta}$ by drawing new values in each simulation from a multivariate normal distribution (the asymptotic distribution of $\hat{\beta}$; step [3a]). We compared observed (black solid lines) and predicted distributions (gray bands representing 95% simulation envelopes) of elevation and precipitation (Fig. 4) and temperature (Fig. 5) at the presence locations. We also overlaid distributions of elevation, precipitation, and temperature at the background locations, f^a (red dashed lines; Fig. 4, Fig. 5). Note that the distributions of elevation and precipitation at the presence locations (solid black lines) were shifted to the right and left, respectively, relative to the background distributions of these covariates (red dashed lines) (Fig. 4). These results reaffirm that this species tends to be found at locations with higher elevations and lower levels of precipitation. In the second example, the distribution of temperature at the used locations was also shifted to the right relative to the background distribution (Fig. 5). In addition, the used distribution was much more peaked compared to the background distribution of temperature, which suggests that this species prefers a more narrow range of temperatures than represented by the background locations.

In the first example, the UHC plots provided evidence that the correct model with both elevation and precipitation was well-calibrated across all three data-generating scenarios (Fig. 4C−D, G−H, K−L) because the distributions of elevation and precipitation at the presence locations (solid black lines) fell mostly within the simulation envelopes generated by the fitted model (gray bands). By contrast, the elevation-only model never accurately predicted the distribution of precipitation values at the presence locations (Fig. 4B, F, J). On the other hand, it predicted the distribution of elevation at the presence locations whenever ρ_{x_1,x_2} was the same for both training and test data sets (Fig. 4A, E). Lastly, the elevation-only model failed to predict either the distribution of elevation or precipitation at the presence locations when the correlation between elevation and precipitation differed between the training and test data (Fig. 4I, J). It is worth noting that in the case where $\rho_{x_1,x_2} = -0.3$ for both training and test data sets, the elevation-only model's predictions were well-calibrated (Fig. 1C, Fig. 4E) even though the logistic regression parameter estimate for elevation was too large (0.80, SE = 0.06) relative to the data-generating value (0.5) (Table 1). These latter two results serve as a nice reminder that regression coefficients reflect partial correlations that are influenced by the suite of predictors included in the model, and are not causal effects (Fieberg and Johnson 2015). Furthermore, models may predict well in the presence of collinearity only when the correlation among predictors remains the same in training and test data (see e.g., Dormann et al. 2013).

In the second simulation example, we fit a model with only a linear effect of temperature on the logit scale and another that also included a quadratic term (the

correct model). When the model included only temperature, the coefficient for temperature was too low, but the coefficients were close to the data-generating values of 2 and −1 when both temperature and temperature^2 were included in the model (Table 2). The predicted distribution for temperature was rather broad and similar to the available distribution when only a linear effect of temperature was included in the logistic regression model (Fig. 5A). By contrast, the distribution of temperature values at presence points was rather peaked, with values of $x_1 < -2$ or > 2 rarely used (Fig. 5A). The extreme avoidance of low and high values of temperatures suggests that a quadratic effect of temperature might be needed. When we included the quadratic term for temperature in the logistic regression model, the distribution of temperature values at the observed locations fell within the 95% simulation envelope (Fig. 5B), confirming that this model was well-calibrated.

In summary, UHC plots helped to identify a missing predictor (precipitation) and also the need for a non-linear term (for temperature). It is also noteworthy that the missing predictor was identified in two scenarios where the model appeared well-calibrated when using a traditional presence-background calibration plot (Fig. 1A,C and Fig. 4B,F) (both scenarios involved predictive distributions in cases where ρ_{x_1,x_2} remained the same in training and test data sets).

Evaluating Spatial Predictions and Model Transferability

An important goal of most SDM applications is to predict species distributions in novel landscapes, which requires that models are "transferable" to other sites, environments, and time periods. If we have location data from multiple sites, then

we can evaluate transferability by fitting a model to some sites and then predicting the distribution of locations at the others (Matthiopoulos et al. 2011). UHC plots can then be used to identify areas in space where the model does a poor job of predicting. To accomplish this goal, we can include x and y spatial coordinates in z, the matrix of habitat characteristics we wish to predict at the out-of-sample used locations.

To illustrate this idea, we return to our simulation example where the species distribution was driven by elevation (x_1) and precipitation (x_2), with the probability of selecting locations proportional to $\exp(0.5x_1 - x_2)$. We simulated uniformly distributed x and y spatial coordinates for the presence and background locations associated with two landscapes (a test and a training landscape), allowing the correlation among (x, y) spatial coordinates and the habitat predictors (x_1, x_2) to differ between the two landscapes (Table 3, Fig. 6). We again fit two models to data collected from the training landscape: the first included only elevation and the second included elevation and precipitation (the correct model). We then evaluated how well these models predicted the spatial distribution of presence points in the test landscape by creating UHC plots for the (x, y) spatial coordinates.

The presence locations in the test landscape were largely concentrated in the southeast (large x and small y; Fig. 6). The correct model accurately predicted the distribution of (x, y) spatial coordinates (Fig. 6C, D). By contrast, the model containing only elevation resulted in a predicted distribution that was relatively uniform in space and for which the x- and y-coordinates were not well calibrated (Fig. 6A, B) . This example illustrates how spatial UHC plots could be used to identify missing predictors (e.g., the poor calibration in Fig. 6A, B might lead an

analyst to consider adding precipitation to the model because it follows a SE-NW gradient in the test landscape). These results also have important implications for management. In particular, one should be wary of using the elevation-only model to determine areas to conserve given the model's poor transferability. Lastly, we note that one can use functions in the ENMeval package (Muscarella et al. 2014) to construct UHC plots with spatially-stratified cross-validation in cases where data are available from a single site. We illustrate this approach in a vignette associated with the *uhcplots* package (Fieberg and ArchMiller 2016).

Step-Selection Functions

An alternative way to motivate the IPP likelihood, eq. (1), can help with conceptualizing generalizations of this approach to longitudinal data. With telemetry data, we may consider the distribution of resources or environmental conditions at the used (i.e., presence) points, $f^u(x)$, as being selected from a distribution of values at available (i.e., background) points, $f^a(x)$, with the selection function $w(x\beta) = \exp(x\beta)$ taking us from the distribution of available locations to the distribution of used locations by way of spatial covariates, x, and a set of regression parameters, β (Lele and Keim 2006):

$$(3)\quad f^u(x_i) = \frac{\exp(x_i\beta) f^a(x_i)}{\int \exp(x(s)\beta) f^a(x(s)) ds}$$

If all areas are equally available, $f^a(x(s))$ is uniform in space (and thus, a constant), getting us back to eq. (1) (Aarts et al. 2012). Selection functions have similarly been used to correct for biased sampling procedures (Patil and Rao 1978), to study natural

selection (Manly 1985), and were first introduced in the context of foraging and habitat selection by McDonald et al. (1990); the theory for estimating selection functions is well developed under the label "weighted distributions" (Patil and Rao 1977).

Historically, radio-telemetry studies allowed animals to be located once to several times per day. Telemetry-based SDMs typically assumed these locations could be treated as independent, with parameters estimated by comparing these locations to randomly sampled ("available") sites from within an animal's estimated home range (Fieberg et al. 2010). This approach was often justified by noting that animals had sufficient time to reach any area within their home ranges between successive locations. The advent of Global Positioning System (GPS) data and associated hardware and software now allows researchers to assess habitat use with much finer temporal resolution. As a consequence, however, telemetry locations collected close in time also tend to be close in space, and the only sites available to an animal shortly after one observation are those accessible to the animal from the previous location, within the time step.

Step-selection functions were developed to address these concerns (Fortin et al. 2005, Forester et al. 2009, Avgar et al. 2016). Rather than treat locations as independent and assume a uniform distribution for $f^a(x)$, step-selection functions treat *movements* between locations as independent. Background locations specific to each telemetry location are generated by considering the previous location, the time between successive locations, and the movement characteristics of the study species – in particular, step lengths (distances between consecutive points collected at fixed

temporal intervals) and turn angles (change in bearing between consecutive locations) (Thurfjell et al. 2014, Avgar et al. 2016). Background locations are generated by sampling step lengths and turn angles from their empirical distributions (Fortin et al. 2005) or from appropriate statistical distributions (e.g., exponential or gamma for step length, von Mises for turn angles) (Forester et al. 2009, Avgar et al. 2016). Step lengths and turn angles are then combined with the location at the previous time point to generate possible movement paths, and as a result, distributions of available points that are location-specific. To guard against misspecification of the step length and turn angle distributions (or, alternatively, to estimate parameters in assumed statistical distributions describing these movement characteristics), one can include as covariates various functions of the distance between points and angular deviations from the previous step (Forester et al. 2009, Avgar et al. 2016).

The likelihood for these data is similar to that for the inhomogeneous Poisson process model, except that we now have stratified data (one stratum for each observed location and its associated available locations generated by the random movement paths):

$$(4)\quad L(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\beta}) = \prod_{i=1}^{K} \frac{\exp(\boldsymbol{x}_{i(k)}\boldsymbol{\beta})}{\sum_{j=1}^{n_i}\exp(\boldsymbol{x}_{j(k)}\boldsymbol{\beta})}$$

where K is the number of strata, n_i is the number of locations (used plus available) in stratum i, and $x_{j(k)}$ are the covariates associated with the j^{th} point in the k^{th} stratum (with $x_{i(k)}$ giving the covariates for the used location).

Calibration Plots with Step-Selection Functions

It is unclear how traditional presence-background calibration plots (e.g., Boyce et al. 2002, Johnson et al. 2006, Phillips and Elith 2010) might be adapted to step-selection functions. In particular, it is not clear how we should account for the strata, which contain a fixed number of used locations (usually one). By contrast, UHC plots, can be adapted to step-selection functions with only two minor changes: 1) rather than fit a logistic regression model in step [2], we can fit a conditional logistic regression model; 2) rather than select a simple random sample in step [3c], we can select a stratified random sample (i.e., selecting one point from within each stratum). No other modifications are necessary.

Here, we illustrate the application of UHC plots to step-selection functions fit to moose (*Alces alces*) telemetry data. From 2010-2015, technicians captured 170 adult female moose in northeastern Minnesota. Technicians fitted moose with Iridium GPS radiocollars (VECTRONIC Aerospace GmbH, Berlin, Germany) recording animal locations at 4.25, 2, and 1.065-hour fix rates. For a full description of capturing and deployment protocols see Carstensen et al. (2014). We selected a single animal with data from summer 2013 and summer 2014 and subsampled data collected at higher fix rates to achieve a consistent 4.25-hour fix rate ± 0.25 hours. We excluded fixes within 24 hours of deployment and those with horizontal dilution of precision >10 (Rempel and Rodgers 1997). This left a total of 689 used locations in both 2013 and 2014.

We generated 10 available locations for each used location by randomly selecting 10 step lengths and 10 turn angles to project the animal forward in time from the

previous location (see Street et al. 2016 for full description of data development). We defined resource availability at used and available locations as the proportional cover of four land cover types within a 50 m radius buffer (identified in the National Land Cover Database 2011; Jin et al. 2013): deciduous forest (decid50), mixedwood forest (mixed50), coniferous forest (conif50) and treed wetlands (treedwet50).

We fit three conditional logistic regression models to the moose data using the *clogit* function in the survival package of Program R (R Core Team 2015, Therneau 2015), treating locations from 2013 as training data and locations from 2014 as test data. In the first model, we included decid50, mixed50, conif50, and treedwet50 as explanatory variables. In the second model, we included the same set of predictors, except we dropped mixed50. Lastly, we fit a model containing only mixed50. We also included step length (divided by 1,000 to scale the magnitude of the regression coefficient to that of the land cover classes) in each of the models to accommodate bias introduced by using parametric distributions for generating step-lengths (Forester et al. 2009, Avgar et al. 2016).

In the original step-selection model, the coefficient for conif50 was negative, whereas the coefficients for decid50, mixed50, and treedwet50 were all positive; of these, only the coefficient for mixed50 was statistically significant (Table 4). When we dropped mixed50 from the model, the coefficients in the step-selection function changed drastically; the coefficients for decid50 and treedwet50 even changed sign (Table 4). The coefficients for all of the compositional predictors left in the model were negative (and all statistically significant), which likely reflects the fact that having more of any one of these habitat types within 50 m meant having less of mixed50.

This series of models nicely illustrates some of the challenges involved with modeling compositional data due to multicollinearity among the predictors (Graham 2003, Cade 2015).

To produce UHC plots for these models, we again simulated 1,000 used test data sets, drawing new regression parameters each time from $\mathbf{N}(\hat{\boldsymbol{\beta}}, \mathbf{c\hat{o}v}(\hat{\boldsymbol{\beta}}))$. The UHC plots were similar for all three models, with the distribution of the covariates at the used points in the test data set largely falling within the predicted distributions for each of the explanatory variables (Fig. 7). These plots suggest that the models are well-calibrated, but also that the information about selection can be captured by a single compositional predictor, mixed50 (Fig. 7I–L).

Discussion

The combination and popularity of open source software (Ghisla et al. 2012, R Core Team 2015), remote sensing technologies, and a plethora of modeling approaches has facilitated the application of models linking plant and animal locations to environmental variables. Further, geographic information systems (GIS) make it easy to produce maps depicting predicted distributions for sampled and unsampled areas. But, how good are these models and the maps they produce? Should we trust models to predict distributions in novel environments, particularly when they are constructed by considering a large suite of often multicollinear predictors (Dormann et al. 2013)? These questions are of utmost importance to wildlife managers and conservation biologists, and thus it is not surprising that they have garnered significant attention lately from ecologists working across a wide range of taxa

(Vanreusel et al. 2007, Moreno-Amat et al. 2015, Torres et al. 2015, Duque-Lazo et al. 2016, Huang and Frimpong 2016).

Most popular approaches to fitting species distribution or habitat selection models rely on comparing observed locations of individuals to randomly or systematically selected locations that describe the background distribution or availability of resources or environmental conditions. Frequently, the combined presence-background data are modeled using binary regression models, with $Y_i = 1$ for observed locations and 0 for background locations (Johnson et al. 2006, Fithian and Hastie 2013). This treatment of the data originally led to much concern and confusion among practitioners who recognized that background points (with $Y_i = 0$) might actually be used by the species (e.g., Keating and Cherry 2004). Recent connections between common modeling approaches (e.g., MaxEnt, spatial logistic regression) and inhomogeneous Poisson process models have clarified both the role of the background points (they serve as quadrature points in eq. (1); Warton and Shepherd 2010) and also the interpretation of regression parameters (they describe systematic variation in the log intensity of the Poisson process model; Aarts et al. 2012, Fithian and Hastie 2013, Renner et al. 2015).

As more researchers become aware of these connections, we expect to see a similar paradigm shift in terms of the methods proposed for validating species distribution and habitat selection models. Traditionally, methods for validating species distribution models have mimicked or modified approaches developed for presence-absence data. They have treated the number of presence locations as random, and have focused on how well the models do at predicting whether locations are "used"

or "available". By contrast, UHC plots consider the number of presence locations as fixed, and instead focus on validating a model's ability to predict the characteristics (i.e., the biotic and abiotic factors used to model distribution patterns) at these locations using out-of-sample data. Our simulation examples demonstrated the utility of UHC plots for identifying missing covariates and nonlinearities that should be included in the model as well as how these plots can be used to identify areas in space that are poorly predicted. Our empirical example, based on moose movement data, demonstrated how this approach can accommodate the stratified nature of step-selection functions and, further, how UHC plots can be used to provide insights into the effect of multicollinearity, particularly when considering compositional data. Future work should focus on exploring the use of UHC plots to suggest possible transformations (e.g., log, step functions) or to detect other forms of model misspecification (e.g., the need for interactions). Simulated data are critical to these efforts since they allow one to evaluate model performance in scenarios where the factors driving the underlying species distribution are known (Miller 2014, Leroy et al. 2016).

Recently developed approaches for assessing fit of spatial point process models offer another promising alternative to UHC plots considered here (Baddeley et al. 2005, 2013, Renner et al. 2015). Specifically, one can plot residuals against spatial covariates or smoothed residuals versus spatial location (e.g., Easting, Northing). These types of plots are available in the *spatstat* library of Program R and have a strong theoretical basis (Baddeley et al. 2008). The advantage of the approach we suggest is that it can be applied more generally, as we have demonstrated with fitted logistic regression

models and step-selection functions. The ability to construct simulation envelopes for out-of-sample data is another advantage, especially since most applications of species distribution models consider a large suite of explanatory variables and often allow for considerable model complexity, leading to data-driven models that may be overfit and perform poorly when applied to new data (Giudice et al. 2012, Harrell 2013).

Understanding what motivates animals to move from one location to another, and how the broad-scale patterns of resources and risk affect the distribution of a species in the landscape is of critical importance to the management and conservation of wildlife and plant species. For models of species distributions to be useful, they must be more than shots in the dark. They must be able to make predictions about how a species will respond to new environmental conditions presented at different locations in space and time in the face of anthropogenic landscape change. By comparing model predictions to out-of-sample data, UHC plots can identify important features that are well-predicted and others where improvement is needed. This process can shed light on how best to modify models, provide important insights into factors driving the distribution of species, and ultimately enhance the reliability and generality of conservation and management recommendations.

Acknowledgements

This work was funded by the University of Minnesota-Twin Cities and the Minnesota Environment and Natural Resources Trust Fund. The Minnesota Department of Natural Resources assisted with collaring and monitoring of the

moose. We thank D. Wolfson, G. Sargeant, and three anonymous reviewers for helpful comments on a previous draft.

References

Aarts, G. et al. 2012. Comparative interpretation of count, presence–absence and point methods for species distribution models. - Methods Ecol. Evol. 3: 177–187.

Araújo, M. B. et al. 2005. Validation of species–climate impact models under climate change. - Global Change Biology 11: 1504–1513.

Arthur, S.M. et al. 1996. Assessing habitat selection when availability changes. - Ecology 77: 215–227.

Avgar, T. et al. 2016. Integrated step selection analysis: Bridging the gap between resource selection and animal movement. - Methods Ecol. Evol. 7: 619–630.

Baddeley, A. et al. 2005. Residual analysis for spatial point processes (with discussion). - J. Roy. Stat. Soc. B. 67: 617–666.

Baddeley, A. et al. 2008. Properties of residuals for spatial point processes. - Ann. I. Stat. Math. 60: 627–649.

Baddeley, A. et al. 2010. Spatial logistic regression and change-of-support in Poisson point processes. - Electron. J. Stat. 4: 1151–1201.

Baddeley, A. et al. 2013. Residual diagnostics for covariate effects in spatial point process models. - J. Comput. Graph. Stat. 22: 886–905.

Boyce, M. S. and McDonald, L. L. 1999. Relating populations to habitats using resource selection functions. - Trends Ecol. Evol. 14: 268–272.

Boyce, M. S. et al. 2002. Evaluating resource selection functions. - Ecol. Model. 157: 281–300.

Cade, B. S. 2015. Model averaging and muddled multimodel inferences. - Ecology 96: 2370–2382.

Carstensen, M. et al. 2014. Determining cause-specific mortality in Minnesota's northeast moose population. - In: Summaries of Wildlife Research Findings 2013. Minnesota Department of Natural Resources (MNDNR), pp. 142–152.

Chivers, C. et al. 2014. Validation and calibration of probabilistic predictions in ecology. - Methods Ecol. Evol. 5: 1023–1032.

Clarke, K. A. 2005. The phantom menace: Omitted variable bias in econometric research. - Conflict Manag. Peace Sci. 22: 341–352.

Dormann, C. F. et al. 2013. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. - Ecography 36: 27–46.

Duque-Lazo, J. et al. 2016. Transferability of species distribution models: The case of *Phytophthora cinnamomi* in Southwest Spain and Southwest Australia. - Ecol. Model. 320: 62–70.

Edwards, T. C. et al. 2006. Effects of sample survey design on the accuracy of classification tree models in species distribution models. - Ecol. Model. 199: 132–141.

Elith, J. et al. 2010. The art of modelling range-shifting species. - Methods Ecol. Evol. 1: 330–342.

Elith, J. et al. 2011. A statistical explanation of MaxEnt for ecologists. - Divers. Distrib. 17: 43–57.

Ellner, S. P. et al. 2002. Precision of population viability analysis. - Conserv. Biol. 16: 258–261.

Fawcett, T. 2006. An introduction to ROC analysis. - Pattern Recogn. Lett. 27: 861–874.

Fieberg, J. and Johnson, D. H. 2015. MMI: Multimodel inference or models with management implications? - J. Wildlife Manage. 79: 708–718.

Fieberg, J. and ArchMiller, A. 2016. uhcplots: Used-habitat calibration plots. R package version 0.1.0.

Fieberg, J. et al. 2010. Correlation and studies of habitat selection: Problem, red herring or opportunity? - Phil. Trans. R. Soc. B 365: 2233–2244.

Fieberg, J. R. et al. 2016. R code and output supporting: Species distribution models: Predictive snipers or shots in the dark? Retrieved from the Data Repository for the University of Minnesota. http://doi.org/10.13020/D6T590.

Fielding, A. H. and Bell, J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. - Environ. Conserv. 24: 38–49.

Fithian, W. and Hastie, T. 2013. Finite-sample equivalence in statistical models for presence-only data. - Ann. Appl. Stat. 7: 1917–1939.

Forester, J. D. et al. 2009. Accounting for animal movement in estimation of resource selection functions: Sampling and data analysis. - Ecology 90: 3554–3565.

Fortin, D. et al. 2005. Wolves influence elk movements: Behavior shapes a trophic cascade in Yellowstone National Park. - Ecology 86: 1320–1330.

Ghisla, A. et al. 2012. Species distribution modelling and open source GIS: Why are they still so loosely connected? – In Seppelt, R. et al. (eds), International Environmental Modelling and Software Society (iEMSs) 2012 International Congress on Environmental Modelling and Software. Managing Resources of a Limited Planet: Pathways and Visions under Uncertainty, Sixth Biennial Meeting: 1481–1488, Leipzig, Germany.

Giudice, J. H. et al. 2012. Spending degrees of freedom in a poor economy: A case study of building a sightability model for moose in northeastern Minnesota. - J. Wildlife Manage. 76: 75–87.

Graham, M. H. 2003. Confronting multicollinearity in ecological multiple regression. - Ecology 84: 2809–2815.

Harrell, F. E. 2013. Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis. - Springer Science & Business Media.

Heikkinen, R. K. et al. 2012. Does the interpolation accuracy of species distribution models come at the expense of transferability? - Ecography 35: 276–288.

Hooten, M. B. et al. 2013. Reconciling resource utilization and resource selection functions. - J. Anim. Ecol. 82: 1146–1154.

Huang, J. and Frimpong, E. A. 2016. Limited transferability of stream-fish distribution models among river catchments: Reasons and implications. - Freshwater Biol. 61: 729–744.

Jiménez-Valverde, A. 2012. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. - Global Ecol. Biogeogr. 21: 498–507.

Jiménez-Valverde, A. et al. 2013. Discrimination capacity in species distribution models depends on the representatives of the environmental domain. - Global Ecol. Biogeogr. 22: 508–516.

Jin, S. et al. 2013. A comprehensive change detection method for updating the National Land Cover Database to circa 2011. - Remote Sens. Environ. 132: 159–175.

Johnson, D. H. 1980. The comparison of usage and availability measurements for evaluating resource preference. - Ecology 61: 65–71.

Johnson, D. H. 2008. In defense of indices: The case of bird surveys. - J. Wildlife Manage. 72: 857–868.

Johnson, C. J. et al. 2006. Resource selection functions based on use-availability data: Theoretical motivation and evaluation methods. - J. Wildlife Manage. 70: 347–357.

Keating, K. A. and Cherry, S. 2004. Use and interpretation of logistic regression in habitat-selection studies. - J. Wildlife Manage. 68: 774–789.

Kutner, M. H. et al. 2005. Applied Linear Statistical Models. - McGraw-Hill Irwin New York.

Lele, S. R. and Keim, J. L. 2006. Weighted distributions and estimation of resource selection probability functions. - Ecology 87: 3021–3028.

Leroy, B. et al. 2016. Virtualspecies, an R package to generate virtual species distributions. - Ecography 39: 599–607.

Manly, B. F. J. 1985. The Statistics of Natural Selection. – Chapman and Hall.

Manly, B. et al. 2002. Resource selection by animals: Statistical design and analysis for field studies. - Springer Science & Business Media.

Matthiopoulos, J. et al. 2011. Generalized functional responses for species distributions. - Ecology 92: 583–589.

Matthiopoulos, J. et al. 2015. Establishing the link between habitat-selection and animal population dynamics. - Ecol. Monogr. 85: 413–436.

Mayor, S. J. et al. 2009. Habitat selection at multiple scales. - Ecoscience 16: 238–247.

McDonald, L. L. et al. 1990. Analyzing foraging and habitat use through selection functions. – In Morrison et al. (eds.), Studies in Avian Biology, Cooper Ornithological Society, pp. 325–331.

Meyer, C. B. and Thuiller, W. 2006. Accuracy of resource selection functions across spatial scales. - Divers. Distrib. 12: 288–297.

Miller, J. A. 2014. Virtual species distribution models using simulated data to evaluate aspects of model performance. - Prog. in Phys. Geogr. 38: 117–128.

Millspaugh, J. J. et al. 2006. Analysis of resource selection using utilization distributions. - J. Wildlife Manage. 70: 384–395.

Moreno-Amat, E. et al. 2015. Impact of model complexity on cross-temporal transferability in Maxent species distribution models: An assessment using paleobotanical data. - Ecol. Model. 312: 308–317.

Moya-Laraño, J. and Corcobado, G. 2008. Plotting partial correlation and regression in ecological studies. - Web Ecology 8: 35–46.

Muscarella, R. et al. 2014. ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. - Methods Ecol. Evol. 5: 1198–1205.

Naimi, B. and Araújo, M. B. 2016. Sdm: A reproducible and extensible R platform for species distribution modelling. - Ecography in press.

Patil, G.P. and Rao, C. R.. 1977. Weighted distributions: a survey of their applications. - In: Krishnaiah, P. R. (ed.), Applications of Statistics, North Holland Publishing Company, pp. 383–405.

Patil, G.P. and Rao, C. R., 1978. Weighted distributions and size-biased sampling with applications to wildlife populations and human families. - Biometrics, 34: 179-189.

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

Pearce, J. L. and Boyce, M. S. 2006. Modelling distribution and abundance with presence-only data. - J. Appl. Ecol. 43: 405–412.

Phillips, S. J. and Elith, J. 2010. POC plots: Calibrating species distribution models with presence-only data. - Ecology 91: 2476–2484.

R Core Team 2015. R: A language and environment for statistical computing. Vienna, Austria; - URL http://www.R-project.org.

Rempel, R. S. and Rodgers, A. R. 1997. Effects of differential correction on accuracy of a GPS animal location system. - J. Wildlife Manage.: 525–530.

Renner, I. W. and Warton, D. I. 2013. Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. - Biometrics 69: 274–281.

Renner, I. W. et al. 2015. Point process models for presence-only analysis. - Methods Ecol. Evol. 6: 366–379.

Skov, H. et al. 2016. Real-time species distribution models for conservation and management of natural resources in marine environments. - MEPS 542: 221–234.

Steyerberg, E. W. et al. 2010. Assessing the performance of prediction models: A framework for some traditional and novel measures. - Epidemiology 21: 128–138.

Street, G. M. et al. 2016. Habitat functional response mitigates reduced foraging opportunity: Implications for animal fitness and space use. - Landscape Ecol. 31: 1939–1953.

Therneau, T. 2015. A package for survival analysis in s. R package version 2.38. - URL http://CRAN. R-project. org/package= survival.

Thurfjell, H. et al. 2014. Applications of step-selection functions in ecology and conservation. - Movement Ecol. 2: 1–12.

Torres, L. G. et al. 2015. Poor transferability of species distribution models for a pelagic predator, the Grey Petrel, indicates contrasting habitat preferences across ocean basins. - PloS one 10: e0120014.

Vanreusel, W. et al. 2007. Transferability of species distribution models: A functional habitat approach for two regionally threatened butterflies. - Conserv. Biol. 21: 201–212.

Wand, M. P. and Jones, M. C. 1994. Kernel smoothing. - Crc Press.

Warton, D. I. and Shepherd, L. C. 2010. Poisson point process models solve the “pseudo-absence problem” for presence-only data in ecology. - Ann. Appl. Stat. 4: 1383–1402.

Wenger, S. J. and Olden, J. D. 2012. Assessing transferability of ecological models: An underappreciated aspect of statistical validation. - Methods Ecol. Evol. 3: 260–267.

Table Legends

Table 1. Estimated regression parameters ($\hat{\boldsymbol{\beta}}$) and their standard errors (SE) for logistic regression models fit to training data in the first cross-sectional data simulation. The marginal distribution of elevation (x_1) and precipitation (x_2) on the landscape was given by a multivariate normal distribution with mean vector = (0,0), and var(x_1) = var(x_2) = 4. We considered three different data-generating scenarios in which we varied cor(x_1, x_2) (= 0, -0.3, or 0.3). The true species distribution was proportional to exp($0.5x_1 - x_2$).

	$y \sim$ elev		$y \sim$ elev + precip			
$\mathbf{cor}(x_1, x_2)$	$\hat{\beta}_{x_1}$	SE	$\hat{\beta}_{x_1}$	SE	$\hat{\beta}_{x_2}$	SE
0.00	0.42	0.05	0.42	0.06	-1.04	0.07
-0.30	0.80	0.06	0.52	0.06	-0.99	0.07
0.30	0.27	0.05	0.57	0.06	-0.97	0.06

Table 2. Estimated regression parameters ($\hat{\boldsymbol{\beta}}$) and their standard errors (SE) for logistic regression models fit to training data in the second cross-sectional data simulation. The marginal distribution of x_3 on the landscape, $f^a(x_3)$, was Normal: $f^a(x_3) = \mathbf{N}(\mathbf{0}, \mathbf{4})$. The relative probability of use (or presence) was proportional to $\exp(2x_3 - x_3^2)$.

Model	$\hat{\beta}_{x_3}$	SE	$\hat{\beta}_{x_3^2}$	SE
$y \sim x_3$	**0.24**	**0.05**		
$y \sim x_3 + x_3^2$	**2.21**	**0.35**	**-1.05**	**0.17**

Table 3. Correlation among spatial coordinates (x, y) and habitat covariates in training and test data in the simulation to evaluate areas in space where the model predicts poorly. The marginal distribution of elevation (x_1) and precipitation (x_2) on the landscape was given by a multivariate normal distribution with mean vector = (0,0), and var(x_1) = var(x_2) = 4. The true species distribution was proportional to $\exp(0.5x_1 - x_2)$.

	Correlation	
Variables	Training data	Test data
x_1, x_2	0.33	0.29
x-coordinate, x_1	0.68	0.57
x-coordinate, x_2	0.33	-0.29
y-coordinate, x_1	0.35	-0.30
y-coordinate, x_2	0.67	0.57

Table 4. Parameter estimates (SE) from step-selection functions fit to moose (*Alces alces*) data in Minnesota using conditional logistic regression. Covariates measured the proportional cover of 4 land cover types within a 50 m radius buffer: deciduous forest (decid50), mixedwood forest (mixed50), coniferous forest (conif50), and treed wetlands (treedwet50). We also included step length (divided by 1,000 to scale the magnitude of the regression coefficient to that of the land cover classes; step) to accommodate bias introduced by using parametric distributions for generating step-lengths.

	Model		
Variable	**(1)**	**(2)**	**(3)**
decid50	**0.49**	**-0.60**	
	(0.33)	**(0.19)**	
mixed50	**1.38**		**1.03**
	(0.24)		**(0.16)**
conif50	**-0.30**	**-1.37**	
	(0.38)	**(0.27)**	
treedwet50	**0.40**	**-0.70**	
	(0.31)	**(0.16)**	
step	**-6.33**	**-6.44**	**-6.39**
	(0.25)	**(0.25)**	**(0.25)**

Figure Legends

Figure 1 Presence-background binned calibration plots using the method outlined in Johnson et al. (2006) applied to simulated data for a species whose distribution was driven by elevation (x_1) and precipitation (x_2). The marginal distribution of x_1 and x_2 on the landscape, $f^a(x_1, x_2)$, was Normal: $f^a(x_1, x_2) = \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$. We considered three different data-generating scenarios in which we set var(x_1) = var(x_2) = 4, but varied cor(x_1, x_2) = ρ_{x_1,x_2} (represented by separate rows of panels). The relative probability of use (or presence) was proportional to $\exp(0.5x_1 - x_2)$. Panels depict observed versus expected numbers of presence locations within 10 bins formed using estimated relative probabilities of selection, $w(x^{test}\hat{\boldsymbol{\beta}}^{train}) = \exp(x^{test}\hat{\boldsymbol{\beta}}^{train})$, where x^{test} is a matrix of covariates in the test data set and $\hat{\boldsymbol{\beta}}^{train}$ is a vector of regression parameter estimates obtained by fitting one of two logistic regression models to the training data (the two models are represented by the different columns). Overlaid is a regression line (black line with shaded 95% confidence intervals) relating observed and expected numbers of presence locations in each bin. A well-calibrated model should closely follow the 1:1 line (dashed line).

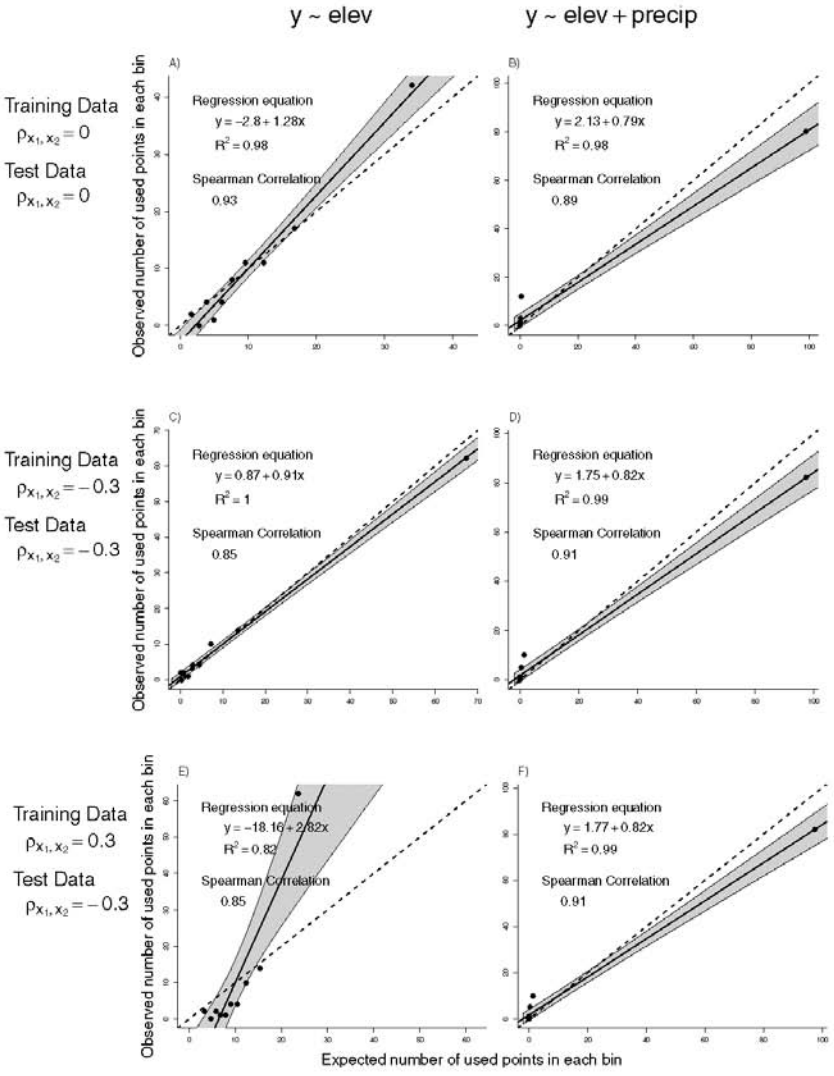

Figure 2 Presence-background binned calibration plots using the method outlined in Johnson et al. (2006) applied to simulated data for a species whose distribution was driven by temperature (x_3) and temperature^2. The marginal distribution of x_3 on the landscape, $f^a(x_3)$, was Normal: $f^a(x_3) = \mathbf{N}(\mathbf{0}, \mathbf{4})$. The relative probability of use (or presence) was proportional to $\exp(2x_3 - x_3^2)$. Panels depict observed versus expected numbers of presence locations within 10 bins formed using estimated relative probabilities of selection, $w(x^{test}\hat{\boldsymbol{\beta}}^{train}) = \exp(x^{test}\hat{\boldsymbol{\beta}}^{train})$, where x^{test} is a matrix of covariates in the test data set and $\hat{\boldsymbol{\beta}}^{train}$ is a vector of regression parameter estimates obtained by fitting one of two logistic regression models to the training data (the two models are represented by the different columns). Overlaid is a regression line (black line with shaded 95% confidence intervals) relating observed and expected numbers of presence locations. A well-calibrated model should closely follow the 1:1 line (dashed line).

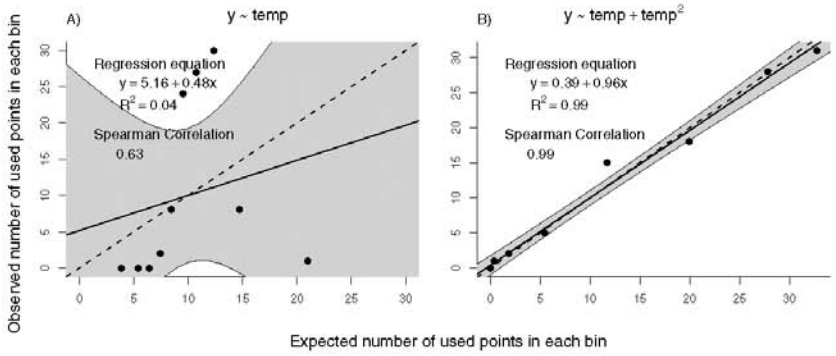

Figure 3 Steps for producing a Used-Habitat Calibration Plot. Step 0: Split the data into test and training data sets (used points are shown in blue, available points in red). Step 1: Summarize the distribution of the explanatory variables (here precipitation and elevation) at the presence points (solid black lines/density plots) and background points (red dashed lines/density plots) in the test data set, $f^u(x)$ and $f^a(x)$, respectively. Step 2: Fit a model to the training data set, storing $\hat{\beta}$ and its uncertainty ($\hat{\text{cov}}(\hat{\beta})$). In this example, the distribution of locations is driven by elevation and precipitation, but only elevation has been included in the model. Step 3: Do the following M times (with loop index i): (a) To account for parameter uncertainty, select new β parameter values, β^i, from the joint parameter distribution describing the uncertainty in $\hat{\beta}$; (b) Estimate $w(x^{test}\beta^i) = e^{x^{test}\beta^i}$ for the test data; (c) Select a simple random (cross-sectional) or stratified random (step-selection function) sample of n_u^{test} observations from the combined (use and available) test data, with probabilities of selection proportional to $w(x^{test}\beta^i)$ from step [3b]; (d) Summarize the predicted distribution of elevation and precipitation using the points chosen in step [3c], $\hat{f}^u(x)_i$. Step 4: Compare the observed distribution of covariate values at the used points, $f^u(x)$ from step [1], to the predicted distribution of these characteristics, $\hat{f}^u(x)_i$ across the M simulations. One option is to overlay $f^u(x)$ on a 95% simulation envelope constructed using the $\hat{f}^u(x)_i$ (gray bands). Step 5: Reevaluate or modify the model as necessary. In the above example, the UHC plots would suggest that we should include precipitation in the model.

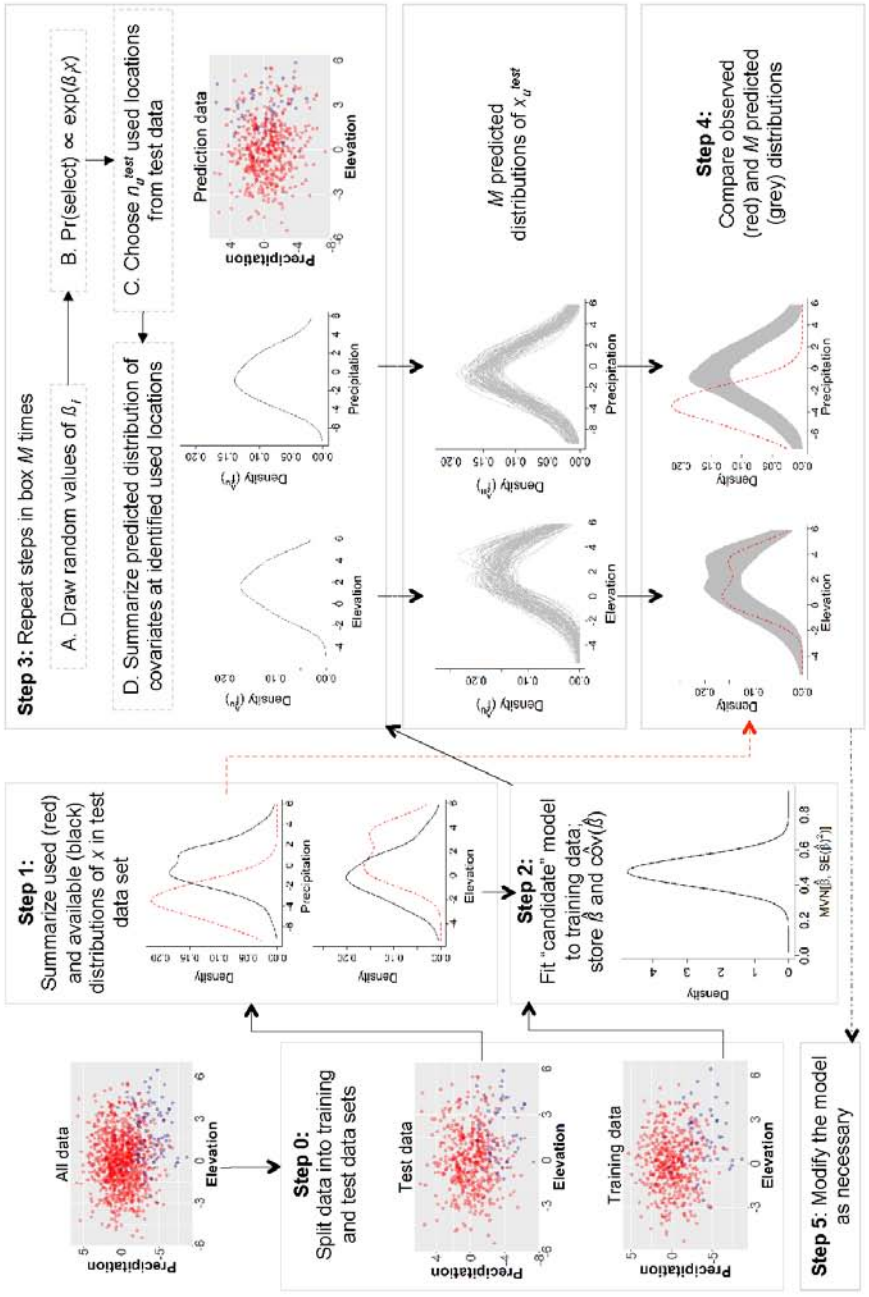

Figure 4 Used-Habitat Calibration (UHC) plots for the first simulation example where the species distribution was driven by elevation (x_1) and precipitation (x_2). The marginal distribution of x_1 and x_2 on the landscape, $f^a(x_1, x_2)$ (red dashed lines), was Normal: $f^a(x_1, x_2) = \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$. We considered three different data-generating scenarios in which we set var(x_1) = var(x_2) = 4, but varied cor(x_1, x_2) = ρ_{x_1,x_2} (represented by separate rows of panels). The relative probability of use (or presence) was proportional to $\exp(0.5x_1 - x_2)$. The observed distribution of elevation and precipitation at the presence (i.e., used) points in the test data set is given by the solid black lines, with a 95% simulation envelope for these distributions given by the gray bands. Predictive distributions were formed using one of two models fit to training data, a model with elevation only (left two columns) or elevation and precipitation (the correct model; right two columns). A model is well-calibrated if the observed distributions (solid black lines) fall within the simulation envelopes.

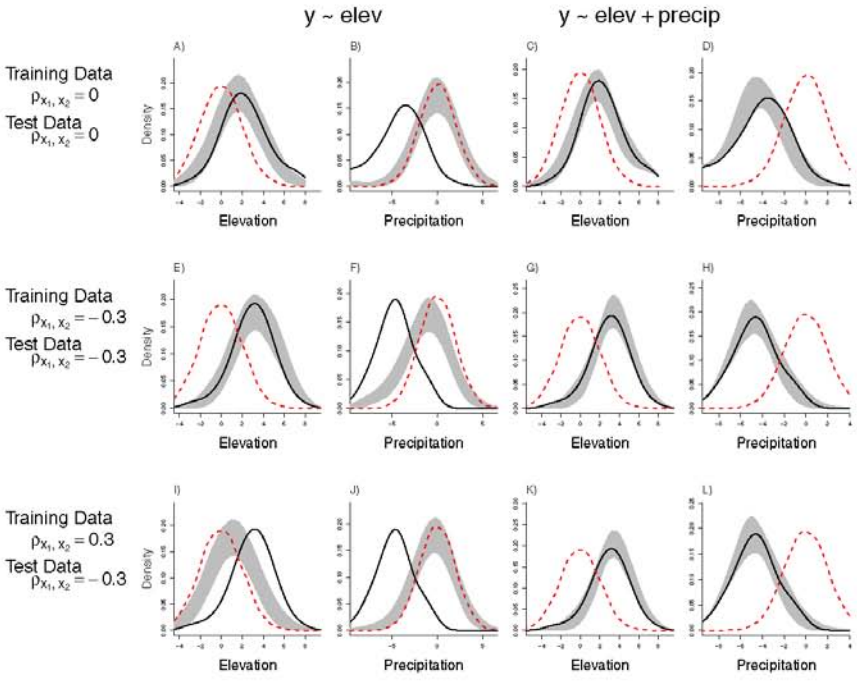

Figure 5 Used-Habitat Calibration (UHC) plots for the second simulation example where the species distribution was driven by temperature (x_3). The marginal distribution of x_3 on the landscape, $f^a(x_3)$ (red dashed lines), was Normal: $f^a(x_3) = \mathrm{N}(0, 4)$. The relative probability of use (or presence) was proportional to $\exp(2x_3 - x_3^2)$. The observed distribution of temperature at the presence points in the test data set is given by the solid black lines, with a 95% simulation envelope for these distributions given by the gray bands. Predictive distributions were formed using one of two models fit to training data, a model with temperature (linear term only; Panel A) or temperature and temperature^2 (the correct model; Panel B). A model is well-calibrated if the observed distributions (solid black lines) fall within the simulation envelopes.

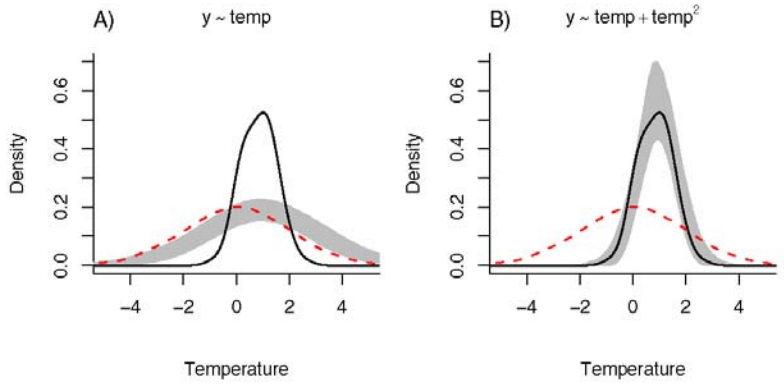

Figure 6 Used-Habitat Calibration (UHC) plots for spatial coordinates (x, y). The species distribution was driven by elevation (x_1) and precipitation (x_2). The marginal distribution of x_1 and x_2 on the landscape, $f^a(x_1, x_2)$ (red dashed lines), was Normal: $f^a(x_1, x_2) = \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$. The relative probability of use (or presence) was proportional to $\exp(0.5x_1 - x_2)$. Top panels depict the background distribution of elevation and precipitation in the training and test data landscapes, with presence points overlaid in yellow and black triangles. In the bottom panels, the observed distribution of elevation and precipitation at the presence points in the test data set is given by the solid black lines, with a 95% simulation envelope for these distributions given by the gray bands. Predictive distributions were formed using one of two models fit to training data, a model with elevation only (panels A and B) or elevation and precipitation (the correct model; panels C and D). A model is well-calibrated if the observed distributions (solid black lines) fall within the simulation envelopes.

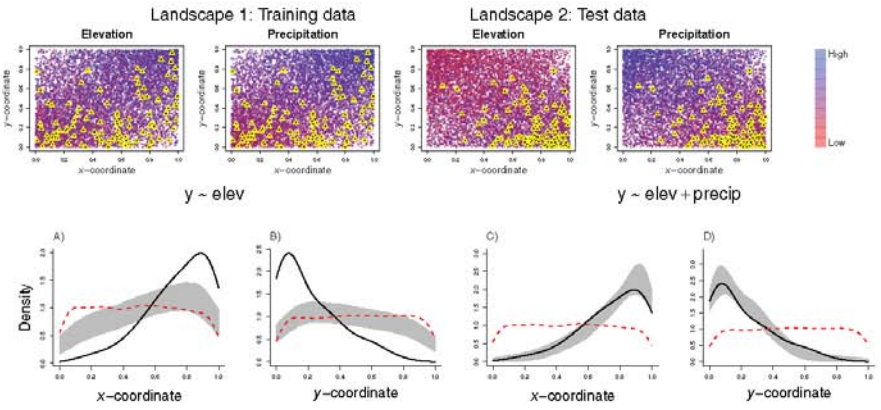

Figure 7 Used-Habitat Calibration Plots for step-selection models fit to moose (*Alces alces*) data in Minnesota. We considered three different models (represented by the three rows of panels), each containing a different subset of covariates (as indicated above each row of panels). Covariates in the models measured proportional coverage of deciduous forest (decid50), mixedwood forest (mixed50), conifer forest (conif50), and treed wetland (treedwt50) within a 50 m buffer of each location. We also included step length (divided by 1,000 to scale the magnitude of the regression coefficient to that of the land cover classes; step) to accommodate bias introduced by using parametric distributions for generating step-lengths. Panels depict the distribution of available and used locations in the test data set (red dashed and solid black lines, respectively), along with 95% simulation envelopes for the predicted distribution of these habitat covariates at the used locations from the fitted step-selection functions. A model is well-calibrated if the observed distributions (solid black lines) fall within the simulation envelopes.