Ecological Modelling 450 (2021) 109566

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Effects of non-representative sampling design on multi-scale habitat models: flammulated owls in the Rocky Mountains.

Luca Chiaverini [a,*], Ho Yi Wan [b], Beth Hahn [c], Amy Cilimburg [d], Tzeidle N. Wasserman [e], Samuel A. Cushman [a,f]

[a] Wildlife Conservation Research Unit, Department of Zoology, University of Oxford, The Recanati-Kaplan Centre, Tubney House, Tubney, Oxon OX13 5QL, UK
[b] Department of Wildlife, Humboldt State University, Arcata, CA, USA
[c] Aldo Leopold Wilderness Research Institute, Missoula, MT, USA
[d] Climate Smart Missoula, Missoula, MT, USA
[e] Ecological Restoration Institute, Northern Arizona University, Flagstaff, AZ, USA
[f] Rocky Mountain Research Station, United States Forest Service, Flagstaff, AZ, USA

ARTICLE INFO

Key words:
Flammulated owl
habitat selection
over-dispersion
sampling bias
simulation
species distribution modelling

ABSTRACT

Sampling bias and autocorrelation can lead to erroneous estimates of habitat selection, model overfitting and elevated omission rates. We developed a multi-scale habitat suitability model of the flammulated owl (*Psiloscops flammeolus*) in the Northern Rocky Mountains based on extensive but spatially clustered survey data, and then used simulations to evaluate the effects of spatially non-representative and spatially representative sampling strategies on model performance and predictions. Our hypothesis was that models trained with spatially non-representative simulated datasets would suffer from bias in parameter estimates, and would show lower predictive performance. The models trained with the spatially representative simulated datasets greatly outperformed the models trained with the spatially non-representative simulated datasets judged on standard metrics of model performance. However, the spatially non-representative models produced superior predictions based on their ability to identify the correct spatial scales, covariates, signs and magnitudes of the species-environment relationships, when compared to the spatially representative models. Thus, it is likely that representative spatial sampling across a broad range of environmental gradients also resulted in over-dispersion of sampling data, with a higher proportion of samples falling in areas of low probability of presence, leading to lower ability to resolve the relationships between species presence-absence and environmental covariates. In contrast, the spatially non-representative sampling, by concentrating sampling along environmental gradients that are characterized by higher probability of presence of the modelled species, produced predictions that, while seeming to be weaker based on standard measures of model performance (e.g., AUC, Kappa, PCC), greatly outperformed the spatially representative models based on measures of true model prediction (e.g., correctly describing the actual spatial scales, direction and strength of species-environment relationships). Further work using simulation approaches is warranted to more fully evaluate the ability of species distribution modelling techniques to correctly identify scales, driving covariates, signs and magnitudes of relationships between species presence-absence patterns, and environmental covariates.

1. Introduction

Species distribution models are a powerful tool for ecological research and biodiversity conservation, but they may be uninformative or misleading if they fail to identify the relevant factors driving species' habitat selection (Pliscoff et al., 2014; Williams et al., 2012). Moreover, there is a longstanding recognition that species-environment relationships occur across a range of spatial scales (Levin, 1992; Wiens, 1989), and assessing environmental factors at a single scale often produces biased estimates or weaker predictive capacity of models (Mateo-Sánchez et al., 2014; Shirk et al., 2014). However, relatively few habitat suitability or species distribution modelling studies have rigorously addressed spatial scale issues and even fewer have applied multi-scale optimization to reliably describe scale dependencies

* Corresponding author. Wildlife Conservation Research Unit, Department of Zoology, University of Oxford, Oxon, UK
E-mail address: luca.chiaverini@zoo.ox.ac.uk (L. Chiaverini).

https://doi.org/10.1016/j.ecolmodel.2021.109566
Received 27 May 2020; Received in revised form 25 February 2021; Accepted 12 April 2021
Available online 23 April 2021

(McGarigal et al., 2016).

Another fundamental, yet less studied, issue is how the spatial pattern and representativeness of sampling strategies affect species distribution models. Sampling bias and autocorrelation can lead to biased estimates of habitat selection, model overfitting and elevated omission rates, and can falsely inflate AUC values (Kramer-Schadt et al., 2013; van Proosdij et al., 2016). Several approaches have been proposed to mitigate the effects of sampling bias in presence-only models, including spatial filtering (Kramer-Schadt et al., 2013), cluster approaches (Fourcade et al., 2014; Varela et al., 2014), Gaussian kernels (Vergara et al., 2016) and background manipulation (Kramer-Schadt et al., 2013; Merow et al., 2013; Phillips et al., 2009). However, considerably less attention has been paid to sampling bias in presence-absence models. A main assumption on which these models rely is that data are representative, independent and evenly distributed over the study area (Guisan and Zimmermann, 2000; Hirzel and Guisan, 2002). Effective sampling strategies should be designed to identify those environmental gradients that are most influential to habitat selection by species (Mohler, 1983; Wessels et al., 1998), rather than being biased towards environmental factors unrepresentative of the spectrum of conditions in which the species occur. Violation of these assumptions can lead to biased estimations of the species-environment relationships and to poorly predictive models (Hirzel and Guisan, 2002).

Randomly stratified strategies are often considered the ideal approach to sample species occurrences from a subset of the entire population (Rathbun and Gerritsen, 2001). However, randomized surveys are rare because they are often logistically difficult and expensive. Surveys of rare and cryptic species are also seldom implemented in spatially and ecologically representative ways, as they mainly focus on areas where the species are already expected to occur. Therefore, models often rely on incomplete and spatially biased datasets, especially towards areas considered *a priori* suitable for the species, or that are more accessible (Kadmon et al., 2004).

Our first goal was to produce a multi-scale habitat suitability model of the flammulated owl (*Psiloscops flammeolus*) in the Northern Rocky Mountains from a dataset collected by the United States Forest Service. The dataset showed moderate spatial clustering and spatial unrepresentativeness, since it was collected by surveying along forest roads and paths, instead of evenly sampling across the study area. Therefore, the dataset does not provide an unbiased representation of all ecological conditions across the study area. Relatively little is known about the distribution of many forest owl species, due to their nocturnal habits, cryptic behaviour and, in some cases, rarity (Johnson et al., 1981). Anthropogenic disturbance can negatively affect owl populations (Wisdom et al., 2000), and the flammulated owl is listed as a sensitive species throughout the western United States, and a species of concern in Canada. The species is mostly restricted to forests of commercially valuable tree species (e.g., ponderosa pine and Douglas fir), where it nests in cavities often associated with mature forest stands (McCallum, 1994a). Furthermore, ecological knowledge on this species is limited; there are relatively few rigorous studies of its habitat ecology and distribution, and most publications are anecdotal accounts.

Our second goal was to evaluate how the spatial sampling bias of the dataset potentially affected the results of the habitat suitability model. We conducted a simulation experiment in which we stipulated the predicted probability model produced with empirical data to reflect the actual probability of occurrence pattern. Then, we simulated new presence-absence datasets, reflecting the probability of presence of the empirical model. The new datasets were produced by (1) sampling the same locations surveyed in the United States Forest Service design, therefore introducing the same spatial bias and unrepresentativeness that affected the survey, and (2) randomly distributing the same number of United States Forest Service survey locations across the study area, producing a spatially representative dataset in which all the environmental conditions characterizing the study area were equitably represented. We then refitted models using the simulated datasets and evaluated the effects of the spatial bias on models' performances and predictions. Our hypothesis was that models trained with spatially non-representative datasets would suffer from bias in parameter estimates, and would show lower predictive performance.

2. Materials and methods

2.1. Study area

The study area consists of western Montana, northern Idaho and a small portion of north-western Wyoming, encompassing over 31.8 million hectares of the United States Northern Rocky Mountains (Figure 1). Public lands managed by United States federal agencies (e.g., Forest Service, Fish and Wildlife Service, Bureau of Land Management, National Park Service, etc.) comprise approximately 45% of total study area. Private lands and human population are concentrated in large valleys between major mountain ranges. Human population is growing more rapidly in this region than most areas of the United States. The continental divide, following the crest of the Rocky Mountains, separates the maritime climate on the west, with higher precipitation and lower seasonal temperature ranges, from a colder and drier continental climate to the east with greater seasonal temperature extremes. Major vegetation types west of the continental divide include extensive forests, comprised of ponderosa pine in the drier sites, Douglas fir, western larch and lodgepole pine at intermediate sites, subalpine fir and Engelmann spruce in cool wet sites and grand fir, western hemlock and western red cedar in warm moist sites. At the lowest elevations and driest sites, natural grasslands occur. Dominant vegetation east of the continental divide is dominated by mixed grass prairies (e.g., wheatgrass, bluestem and needlegrass), big sagebrush shrublands, steppe (e.g., bluebunch wheatgrass) on lower elevation sites, and forests, dominated by Douglas fir and lodgepole pine, at middle to upper elevations. The mean elevation of the study area is 2,048m, ranging from ~200m in the Nez Perce-Clearwater National Forest in Idaho to ~3,900m at Granite Peak in Montana.

2.2. Presence-absence locations

The United States Forest Service Northern Region collected flammulated owl presence-absence data in 2005 and in 2008 through a survey that included spatial bias, since it was mainly focused in forested areas where the species was expected to potentially occur, and along forest roads and paths. The dataset was collected following a standardized protocol (Fylling et al., 2010): nocturnal surveys were carried out during incubation and brooding periods, between mid-May and late June. Surveyors spent a total of 10 minutes at each survey location, divided into five 2-minute intervals. Two minutes of silent listening at the beginning were followed by four intervals of calling and listening, during which the first 30 seconds were spent broadcasting a pre-recorded standardized flammulated owl call, pointing the caller in each of 4 cardinal directions, and the remaining 90 seconds listening. When an owl was detected, the surveyor recorded the survey location, the bearing using a compass and the distance from the survey point.

For training and validating the distribution model, we used two different presence-absence datasets. To train the model, we used owl data collected in 2005 (n= 2,688; 243 presences, 2,439 absences and 6 locations with missing occurrence information). We removed points with missing coordinates and with missing occurrence information. Since a small number of locations (n= 225) were sampled ~150km away from the main survey area, we removed these points to reduce potential over-dispersion. We also removed locations ≤10km of the edge of the study area to avoid boundary problems in multi-scale analysis. The final training dataset had 2,428 points (242 presences and 2,186 absences). To validate the model, we used the owl data collected in 2008 (n= 1,811; 177 presences and 1,634 absences). We removed points applying the aforementioned criteria, obtaining a final validation

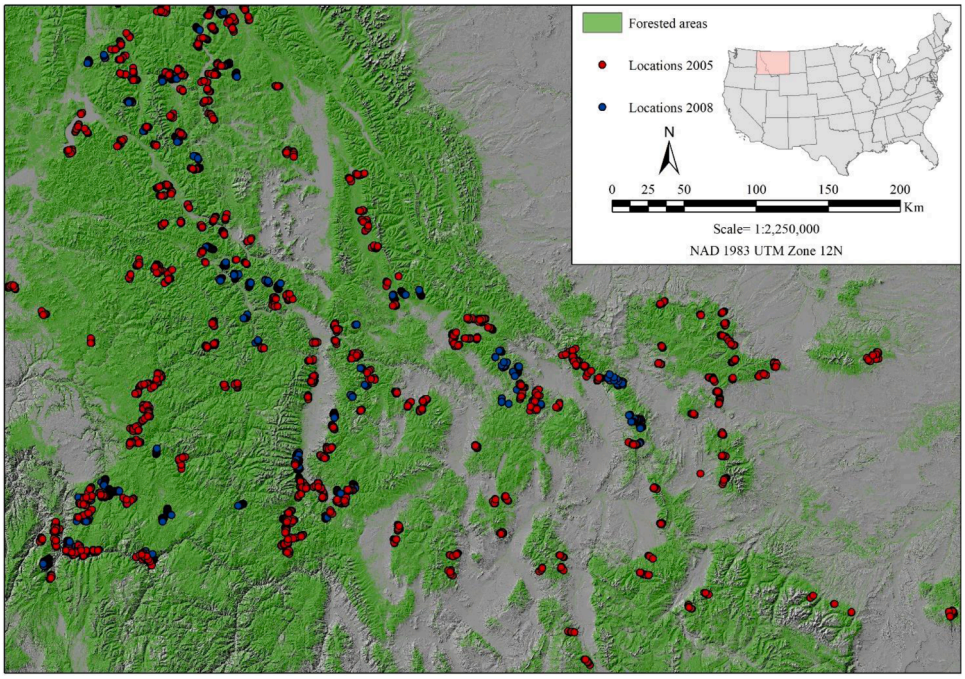

Figure 1. Study area orientation map showing sampling locations.

dataset of 1,809 points (177 presences and 1,632 absences).

2.3. Environmental covariates

We selected a preliminary set of 21 covariates, considered important to habitat selection by flammulated owl (Christie and van Woudenberg, 1997; McCallum, 1994a, b). These included 13 landscape, 5 topographic and 3 climatic covariates. Landscape covariates included: (1) Douglas fir (*Pseudotsuga menziesii*) aggregation index, (2) Douglas fir edge density, (3) Douglas fir mean edge proximity, (4) percent of landscape covered by Douglas fir, (5) ponderosa pine (*Pinus ponderosa*) aggregation index, (6) ponderosa pine edge density, (7) ponderosa pine mean edge proximity, (8) percent of landscape covered by ponderosa pine, (9) forest aggregation index, (10) forest edge density, (11) forest mean edge proximity, (12) percent of landscape covered by forest and (13) tree canopy cover. All landscape covariates were derived from LANDFIRE Existing Vegetation Type layer (LANDFIRE, 2005), with the exception of tree canopy cover, derived from Hansen et al. (2013). To calculate Douglas fir and ponderosa pine metrics, we created 30m resolution binary layers by aggregating all pixels classified as "Interior Douglas-Fir" and "Interior Ponderosa Pine", respectively, in the SAF-SRM Cover Type. We combined all pixels classified as "tree-dominated" in NVCS and created a binary map to calculate forest metrics.

We used FRAGSTATS (McGarigal et al., 2012) to calculate metrics for the Douglas fir, ponderosa pine and forest maps described above. Aggregation index (AI) measures the extent to which patches are aggregated or clumped, by summing, over the landscape elements, the products of the probability that a randomly chosen cell belongs to landscape element *i*, with the conditional probability that, given a cell is of landscape element *i*, one of its neighbouring cells belongs to landscape element *j*. Edge density (ED) measures the density of the edge segments of the corresponding landscape element, by calculating the total length of the edge segments over all the patches of the landscape element *i* and dividing it by the total landscape area. Mean edge proximity (PROX_MN) measures the degree of isolation and fragmentation of the corresponding landscape element, by summing, over all the patches of the landscape element *i* whose edges are within the search radius of the focal patch, each patch size divided by the square of its distance from the focal patch. Percentage of landscape (PLAND) measures the proportional abundance of the corresponding landscape element in the landscape, by summing the areas of all patches of the corresponding landscape element *i* and dividing it by the total landscape area.

Topographic covariates included: (1) elevation, (2) slope, (3) aspect, (4) slope position and (5) roughness. All topographic covariates were derived from the LANDFIRE Digital Elevation Model layer (LANDFIRE, 2005). Slope and aspect were calculated using the DEM Surface Tools (Jenness, 2013) in ArcMap v10.6.1. We applied a cosine transformation to the aspect layer to obtain continuous values ranging from -1 (due south) to +1 (due north), and we classified flat areas to 0. Slope position and roughness layers were derived using the Geomorphometry & Gradient Metrics toolbox (Evans et al., 2014) in ArcMap, by implementing a circular moving window of 90m radius, corresponding to 3 pixels, to the original elevation layer, maintain the high resolution of the source layer.

Climatic covariates included: (1) cumulative annual degree-days (using a 10°C threshold), (2) spring precipitation (defined as February-May precipitations) and (3) summer precipitation (defined as June-October precipitations). All climatic covariates were derived from Recent Years PRISM climate data (PRISM Climate Group, 2016). All covariate layers were resampled at 30m resolution.

2.4. Univariate scaling and multicollinearity analysis

To select the most representative scale for each covariate, we calculated metrics at each sampling location using bandwidths from 100m to 5,000m at 100m interval, for a total of 50 scales (Wan et al., 2017). For landscape covariates, we assessed the aforementioned metrics in FRAGSTATS at different scales, while for topographic and climatic covariates we calculated focal mean at different scales by applying neighbourhood analyses in ArcMap. In all cases, we used a uniform moving window. We then performed a binomial generalized linear model (GLM) using *lme4* package (Bates et al., 2015) in R v3.5.1 (R Core Team, 2018), independently at each scale. We compared models using Akaike's Information Criterion corrected for small sample size (AICc) and the proportion of deviance explained, retaining only the most parsimonious and explanatory scale for each covariate.

We then checked for multicollinearity calculating Pearson's correlation index between each pair of covariates at their best scale. When two covariates were highly correlated ($|r| \geq 0.7$), we dropped the covariate whose univariate GLM showed the greatest AICc. Finally, we calculated variance inflation factor (VIF) among the remaining covariates, ensuring that none of them had VIF≥ 3 (Zuur et al., 2009), which is a conservative threshold ensuring a high degree of independence among predictor variables.

2.5. Multi-scale modelling and spatial autocorrelation

We conducted an all-subsets logistic regression analysis using the *MuMIn* package (Bartoń, 2013) in R v3.5.1 to produce model average coefficient estimates on the final suite of covariates. However, first, we checked for spatial autocorrelation by applying Moran's I test to the model including the full set of covariates, by using the *spdep* package (Bivand and Piras, 2015). We then corrected for residual spatial autocorrelation by adding a spatial autocovariate term (SAC) and forcing it into all subset models.

We ranked the candidate models using AICc and Akaike's model weight (w_i), retaining only those with ΔAICc≤ 2 to represent competing models (Burnham and Anderson, 2002). We calculated the parameter estimates of each covariate by averaging the estimates from the suite of competing models, according to the respective w_i. Using these estimates, we created a map predicting probability of detecting flammulated owl presence across the study area.

2.6. Model performance

To evaluate the explanatory power of the final model, we calculated the proportion of deviance explained, model-averaged parameter estimates and 95% confidence intervals, as well as AIC variable importance. To further inspect the importance of each covariate, we sequentially removed each covariate (with replacement) and evaluated the reduction in deviance explained. We divided that reduction by the total deviance explained by the model, to obtain the percent drop in deviance explained for each covariate. We then evaluated the model's predictive performance by calculating the percent correctly classified (PCC), sensitivity, specificity, Kappa statistics and area under the ROC curve (AUC). We calculated these metrics using an independent dataset, providing more robust measures of predictive performance (Fielding and Bell, 1997). Threshold-dependent measures of model performance (i.e., PCC, sensitivity, specificity and Kappa statistics) were calculated by selecting a cut-off value aimed at maximizing Kappa statistics, and we used the *PresenceAbsence* package (Freeman and Moisen, 2007) to determine the optimal threshold value.

2.7. Simulation analyses

To investigate the effects of the non-random nature of the survey data on the model performance, we used simulated data to compare the effects of spatially representative and non-representative sampling on model prediction and performance. First, we simulated 10 random raster layers with the same extent of the predicted probability map (hereafter *empirical model*), and pixel values uniformly distributed between 0 and 1. We then subtracted these layers from the empirical model, and assigned 0 to cells with negative values and 1 to cells with positive values. We treated 0 and 1 on these binary maps as random representation of potential presence-absence locations, assuming the empirical model to reflect the actual presence pattern of the species (Cushman et al., 2016; Cushman et al., 2017). Then, from each map, we randomly sampled the same number of locations used to train the empirical model (n= 2,428), producing ten different pseudo-presence-absence datasets (hereafter *spatially representative datasets*; Table S1). We also sampled from each map the same sampling locations used to train the empirical model, obtaining ten different pseudo-presence-absence datasets (hereafter *spatially non-representative datasets*; Table S1). We developed a multi-scale model for each simulated dataset using the same modelling framework described for the empirical model.

To compare the models trained with the representative and non-representative datasets, we first explored the results of the univariate scaling analyses to examine whether the best scales were different. Then, we compared the parameter estimates of the covariates, focusing on the signs and the magnitudes. To compare models' predictive performances, we simulated another random raster with the same extent of the empirical model, using the same procedure as described above to produce a binary presence-absence layer. To calculate the predictive performances of the spatially representative models, we randomly sampled the same number of locations used to validate the empirical model (n= 1,809), obtaining a pseudo-presence-absence dataset. To calculate the predictive performances of the spatially non-representative models, we sampled the same sampling locations used to validate the empirical model, obtaining another pseudo-presence-absence dataset. To compare models' predictive performances, we calculated PCC, sensitivity, specificity, Kappa statistics and AUC for the 20 simulations. A comprehensive workflow diagram of the methodologies applied is provided in Figure 2.

3. Results

3.1. Univariate scaling analysis and spatial autocorrelation

The scales showing the lowest AICc always corresponded to the scales with the highest proportion of deviance explained in the univariate scaling analyses (Figure S1). Landscape covariates showed a broad range of optimal scales. The four forest covariates and tree canopy cover were selected at broad scales (4,900m-5,000m). Aggregation index and edge density of Douglas fir and ponderosa pine were selected at mid to broad scales (2,600m-5,000m), while edge proximity was selected at finer scales (1,700m for Douglas fir and 100m for ponderosa pine). Percentage of landscape showed an opposite pattern between Douglas fir and ponderosa pine, being selected at 400m and 4,900m, respectively. Topographic covariates also showed a broad range of optimal scales. Elevation was selected at the smallest scale assessed (100m), while aspect and slope were selected at mid scales (1,000m and 2,300m, respectively). Topographic indexes were selected at mid to broad scales (1,200m for slope position and 4,300m for roughness). All climate covariates were selected at the broadest scale assessed (5,000m). Nine covariates were excluded as a result of the multicollinearity analyses (Table 1).

Moran's I analysis indicated spatial autocorrelation in the flammulated owl data (observed Moran's I= 0.33, $p< 0.001$). Hence, we corrected the multi-scale model by adding a spatial autocovariate term (SAC) into the final model. The correction produced a considerable reduction in spatial autocorrelation (observed Moran's I= -0.05, $p=$ 0.13) (Figure 3).

3.2. Habitat covariates and multi-scale model

The multi-scale model consisted of twelve covariates, excluding intercept and SAC (Figures S2-S13; Table 1). The suite of top models included 10 models (Table 2). After model averaging, most covariates showed AIC variable importance= 1.00, except forest aggregation index (0.82), spring precipitation (0.75), Douglas fir aggregation index (0.51), ponderosa pine edge density (0.25) and elevation (0.06) (Table 3).

Douglas fir percent cover, ponderosa pine aggregation index, ponderosa pine edge density, ponderosa pine edge proximity, slope and slope position were positively related to flammulated owl habitat selection, whereas Douglas fir aggregation index, forest aggregation index, elevation, aspect, spring precipitation and summer precipitation were negatively related to flammulated owl detection (Table 3). Douglas fir percent cover was the most important covariate in the model, showing the greatest drop in model deviance explained (5.41%), followed by

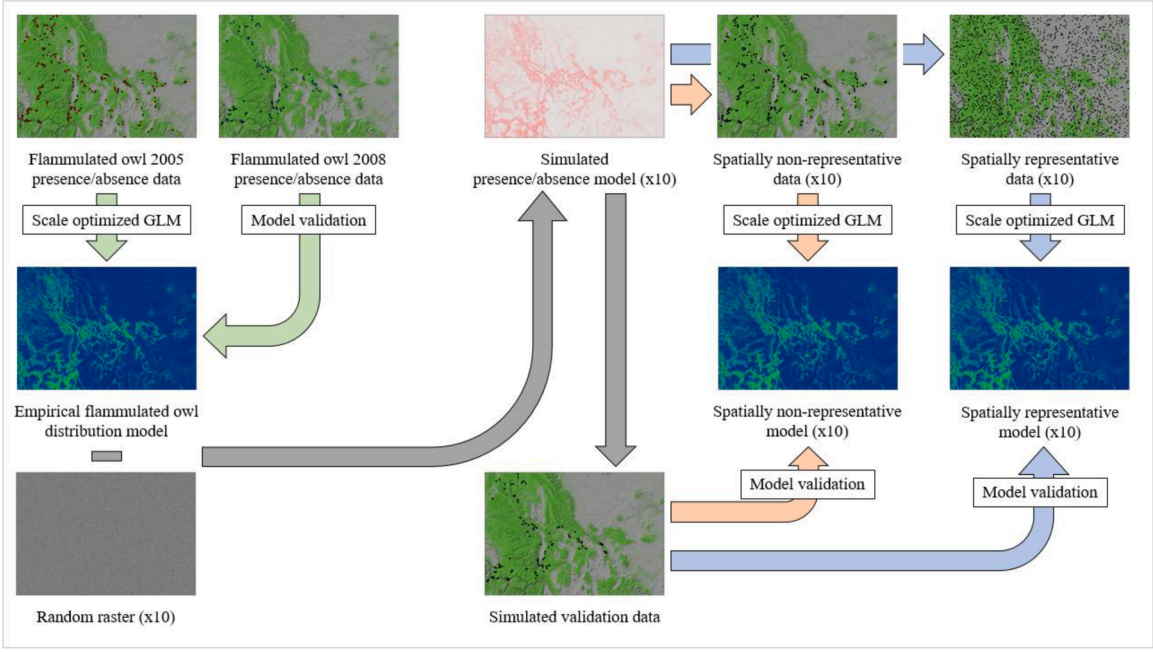

Figure 2. Workflow diagram of the methodologies applied to produce the empirical model and to simulate the spatially representative and the spatially non-representative models.

Table 1
Primary set of covariates selected to model flammulated owl presence, based on the empirical model.

Class	Covariate	Description	Best scale (m)
Landscape	DFc	Douglas fir aggregation index	5,000
	DFed*	Douglas fir edge density	2,600
	DFp*	Douglas fir edge proximity	1,700
	DFpl	Douglas fir percent cover	400
	PPc	Ponderosa pine aggregation index	3,100
	PPed	Ponderosa pine edge density	5,000
	PPp	Ponderosa pine edge proximity	100
	PPpl*	Ponderosa pine percent cover	4,900
	Fc	Forest aggregation index	5,000
	Fed*	Forest edge density	5,000
	Fp*	Forest edge proximity	4,900
	Fpl*	Forest percent cover	5,000
	Han*	Tree canopy cover	5,000
Topographic	DEM	Elevation	100
	S	Slope in degree	2,300
	A	Aspect from -1 to +1	1,000
	SP	Slope position index	1,200
	R*	Roughness index	4,300
Climate	CDD*	Cumulative annual degree-days	5,000
	SpP	Spring (Feb – May) precipitation	5,000
	SuP	Summer (Jun – Oct) precipitation	5,000

* Covariates excluded from the empirical model after the multicollinearity and VIF analyses.

slope position (3.32%), slope (2.33%), Ponderosa pine aggregation index (1.62%) and summer precipitation (1.48%). Elevation was the least important covariate, showing minor drop in model deviance explained (0.03%) (Figure 4; Table 3).

3.3. Model performance

Deviance explained by the final model was 0.29. The optimal threshold for maximizing Kappa statistics was 0.25 (Figure S14; Table 4). We used this as the cut-off for assessing validation metrics. The model correctly classified 75% of the independent validation data and had Kappa= 0.16. The model showed higher specificity than sensitivity

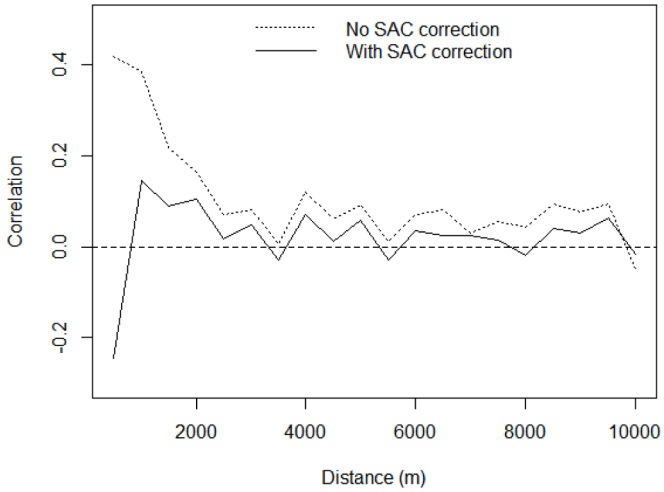

Figure 3. Assessment of the effectiveness of the spatial autocovariate (SAC) approach to reduce spatial autocorrelation of residuals.

(0.78 and 0.49, respectively). AUC was 0.68, indicating moderate discrimination between presence and absence points. The map of the predicted probability of flammulated owl presence is shown in Figure 5.

3.4. Simulation analyses – Single-scale models

The single-scale analyses for the spatially representative simulated datasets showed substantial variation in the most representative scales for each covariate among the iterations (Figure S15; Table 5), and compared to the empirical model (Table S2). Across the iterations, only one covariate always showed the same best scale as the empirical model, and two covariates differed from the empirical model by less than 100m. Eight covariates showed best scales differing between 100m and 1,000m from the empirical model, while 10 covariates showed differences >1,000m.

The single-scale analyses for the spatially non-representative

Table 2
Top multi-scale models selected for the final averaged empirical model, ranked by AICc. Only models with ΔAICc< 2 were selected. Proportion of deviance explained (D2), absolute AICc, ΔAICc and AICc weights (w_i) of each model are provided. SAC represents the spatial autocovariate term.

Models	D2	AICc	ΔAICc	w_i
DFc+DFpl+PPc+PPp+Fc+S+A+SP+SpP+SuP+SAC	0.29	1,146.13	0.00	0.17
DFpl+PPc+PPp+Fc+S+A+SP+SpP+SuP+SAC	0.29	1,146.17	0.04	0.17
DFc+DFpl+PPc+PPp+S+A+SP+SpP+SuP+SAC	0.29	1,147.09	0.96	0.10
DFpl+PPc+PPed+PPp+Fc+S+A+SP+SpP+SuP+SAC	0.29	1,147.19	1.05	0.10
DFc+DFpl+PPc+PPp+Fc+S+A+SP+SuP+SAC	0.29	1,147.20	1.07	0.10
DFpl+PPc+PPp+Fc+S+A+SP+SuP+SAC	0.28	1,147.64	1.51	0.08
DFc+DFpl+PPc+PPed+PPp+Fc+S+A+SP+SpP+SuP+SAC	0.29	1,147.73	1.60	0.08
DFpl+PPc+PPp+S+A+SP+SpP+SuP+SAC	0.28	1,147.77	1.64	0.07
DFpl+PPc+PPed+PPp+Fc+S+A+SP+SuP+SAC	0.29	1,147.91	1.78	0.07
DFc+DFpl+PPc+PPp+Fc+DEM+S+A+SP+SpP+SuP+SAC	0.29	1,148.13	2.00	0.06

Table 3
Representative scale, model averaged parameter coefficients, standard error (SE), variable importance (VI) and percentage of deviance explained by the covariates retained in the empirical model.

Covariate	Scale	Coefficient (± SE)	VI	Deviance explained
Intercept	NA	-3.25E+00 (±1.30E-01)	1.00	NA
Douglas fir aggregation index	5,000	-1.11E-01 (±1.54E-01)	0.51	0.20
Douglas fir percent cover	400	5.40E-01 (±1.13E-01)	1.00	5.41
Ponderosa pine aggregation index	3,100	3.52E-01 (±1.60E-01)	1.00	1.62
Ponderosa pine edge density	5,000	3.33E-02 (±9.45E-02)	0.25	0.06
Ponderosa pine edge proximity	100	1.83E-01 (±6.66E-02)	1.00	1.31
Forest aggregation index	5,000	-1.99E-01 (±1.45E-01)	0.82	0.37
Elevation	100	-3.78E-04 (±2.67E-02)	0.06	0.03
Slope	2,300	3.38E-01 (±1.06E-01)	1.00	2.33
Aspect	1,000	-1.71E-01 (±8.25E-02)	1.00	0.84
Slope position	1,200	3.13E-01 (±8.06E-02)	1.00	3.32
Spring precipitation	5,000	-1.98E-01 (±1.68E-01)	0.75	0.41
Summer precipitation	5,000	-3.11E-01 (±1.25E-01)	1.00	1.48
SAC	NA	7.04E+02 (±5.89E+01)	1.00	NA

datasets showed lower differences in the most representative scales selected among the iterations (Figure S15; Table 5) and compared to the empirical model (Table S2). Five covariates always showed the same best scales as the empirical model, while 3 covariates diverged by less than 100m. Eight covariates showed differences between 100m and 1,000m, and 5 covariates diverged by >1,000m.

3.5. Simulation analyses – Covariates selection and spatial autocorrelation

Douglas fir aggregation index, ponderosa pine aggregation index, aspect, slope position and summer precipitation were the only covariates retained in all the spatially representative iterations. Nine covariates were retained in at least 50% of the iterations, while seven covariates occurred in <50% of the iterations (Table S3). A spatial autocovariate term (SAC) was added when the corrected multi-scale model showed a less significant Moran's I value than the uncorrected model. We added a SAC term in 5 out of 10 models.

Douglas fir aggregation index, Douglas fir percent cover, ponderosa pine aggregation index, ponderosa pine edge proximity, aspect, slope position, spring precipitation and summer precipitation occurred in all the spatially non-representative iterations. Seven covariates were retained in at least 50% of the iterations, while six covariates occurred in <50% of the iterations (Table S3). After checking for spatial autocorrelation, we added a SAC term in 3 out of 10 models.

3.6. Simulation analyses – Multi-scale models

Among the spatially representative models, the covariates showing the biggest differences in the averaged standardized coefficients between the empirical and the simulated model were slope and elevation (0.16 and 0.13, respectively). The covariates showing the smallest differences were slope position and summer precipitation (0.0016 and 0.0125, respectively) (Figure S16; Table 6). Douglas fir aggregation index, ponderosa pine edge density, elevation, slope and spring precipitation's coefficients showed an opposite sign to the empirical model's coefficients in the 0.30%, 0.11%, 0.50%, 0.14% and 0.33% of the iterations, respectively (Table 6).

Among the spatially non-representative models, the covariates showing the biggest differences in the averaged standardized coefficients were ponderosa pine aggregation index and slope (0.18 for both). The covariates showing the smallest differences were Douglas fir percent cover and elevation (0.0014 and 0.0036, respectively) (Figure S16; Table 6). Douglas fir aggregation index and elevation showed an opposite sign to the empirical model's coefficients in the 0.30% and 0.50% of the iterations, respectively (Table 6).

3.7. Simulation analyses – Model performance

Among both the spatially representative and the spatially non-representative models, the covariates showing the widest discrepancies in drop of deviance explained from the empirical model were Douglas fir percent cover and slope position (10.34 and 5.16, respectively, for the spatially representative iterations; 12.78 and 10.99, respectively, for the spatially non-representative iterations) (Table S4). Douglas fir percent cover and slope position were, respectively, the first and the second most important covariates in the empirical model and they represented the most important covariates also in both simulated models. The covariates showing the smallest differences in drop of deviance explained were Douglas fir aggregation index among the spatially representative models (0.39), and elevation among the spatially non-representative models (0.07).

For assessing validation metrics, we used different optimal thresholds to maximize each models' Kappa statistics (Table 7). The simulated spatially representative models correctly classified, on average, 91% of the validation points and had mean Kappa statistics of 0.32. Models revealed, on average, higher specificity than sensitivity (0.95 and 0.42, respectively). Mean AUC was 0.82. Spatially representative models showed much higher apparent predictive performance when compared to the empirical model. All the simulated models showed higher PCC,

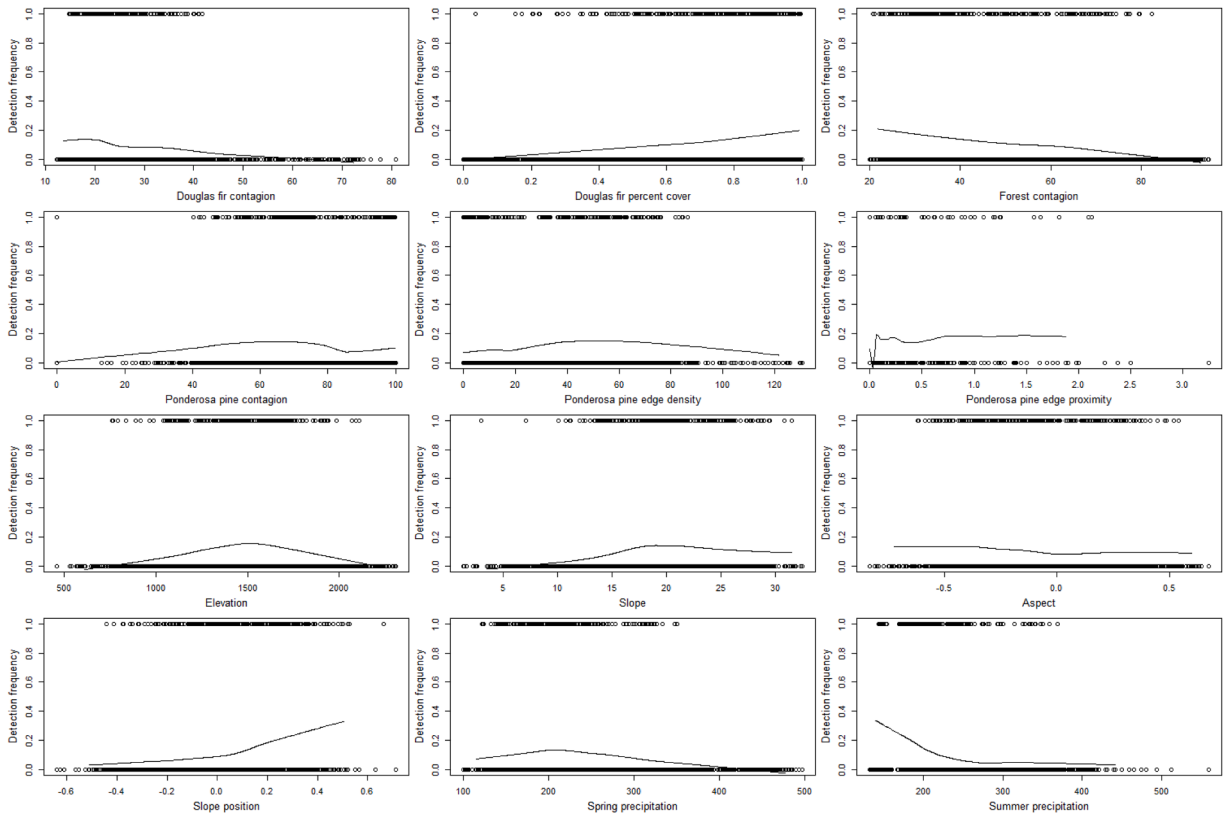

Figure 4. Flammulated owl detection frequencies in response to the covariates retained in the final empirical model.

Table 4
Empirical model performances. Sensitivity represents the number of correctly predicted presence locations divided by the total number of presence locations (true positive fraction). Specificity represents the number of correctly predicted absence locations divided by the total number of absence locations (true negative fraction). Kappa represents the percent improvement over random classification. Area under the curve (AUC) and deviance explained are threshold-independent measures of model performance.

Model	PCC	Kappa	Sensitivity	Specificity	AUC	Deviance explained
Empirical model	0.75	0.16	0.49	0.78	0.68	0.29

Kappa statistics and AUC values than the empirical model.

The simulated spatially non-representative models correctly classified, on average, 70% of the validation points and had mean Kappa statistics of 0.18. Models showed, on average, higher specificity than sensitivity (0.77 and 0.42, respectively). Mean AUC was 0.66. Spatially non-representative models did not show strong variation in predictive performance when compared to the empirical model. PCC, Kappa statistics and AUC values were comparable to the empirical model.

Among the spatially representative iterations, the covariates with the biggest differences in variable importance were ponderosa pine edge density and slope (0.43 and 0.37, respectively). The covariates showing the smallest differences were Douglas fir percent cover, ponderosa pine aggregation index and slope position, which occurred in all the iterations and showed AIC variable importance= 1.00, identically to the empirical model (Table S5).

Among the spatially non-representative models, ponderosa pine edge density and spring precipitation showed the biggest differences in variable importance (0.33 and 0.19, respectively). The covariates with the smallest differences were Douglas fir percent cover, ponderosa pine edge proximity and slope position, showing an AIC variable importance= 1.00, as in the empirical model (Table S5).

4. Discussion

Spatial bias is likely to affect distribution models if the strategy implemented in data collection is based on non-random sampling, potentially leading to poorly predictive inferences (Hirzel and Guisan, 2002). Nevertheless, random sampling is rare for a number of practical and economic reasons. Evenly sampling large study areas can be logistically demanding and very expensive, especially when the focal species are rare and cryptic, or when the habitat is not easily accessible. Hence, many datasets are often collected from relatively easily accessible areas (e.g., roads, paths and rivers), rather than systematic or random sampling. Models produced with biased datasets over-represent certain environmental features reflecting sampling effort rather than the true potential distribution of the species (Kadmon et al., 2004; Phillips et al., 2009), potentially leading to inappropriate management decisions (MacKenzie, 2005; Phillips et al., 2009). We sought to test the hypothesis that models trained with spatially non-representative datasets would suffer from bias in parameter estimates and would show lower predictive performance and variance explained, starting from an empirical model trained with a spatially biased dataset.

4.1. Differences in model performance

The models trained with the spatially representative datasets appeared to greatly outperform the models trained with the spatially non-representative datasets, in relation to all the assessed metrics. The spatially non-representative simulations had much lower PCC than the spatially representative simulations, and comparable to the empirical model. AUC and Kappa statistics also confirm that the spatially representative models greatly outperformed the spatially non-representative ones. Specifically, the former produced AUC values generally

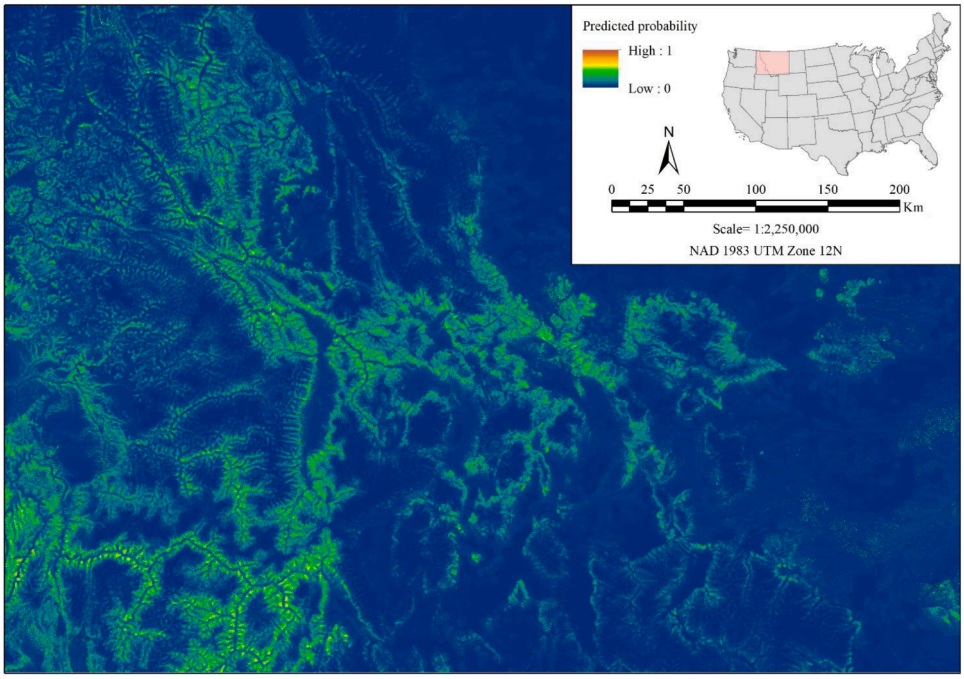

Figure 5. Map of the predicted probability of occurrence for flammulated owl based on the empirical model.

considered to represent strong discrimination between presences and absences, whilst the latter produced AUC associated with poor to fair discrimination. The AUC of the empirical model is comparable to the spatially non-representative one. Moreover, the spatially non-representative models produced Kappa statistics associated with a very low improvement of their classification abilities over a random model, similar to the empirical model. Kappa statistics produced by the spatially representative models indicated better classification ability. These results seem to demonstrate that spatially non-representative datasets lead to poorly predictive distribution models, while unbiased sampling strategies produce stronger models in terms of their ability to correctly discriminate between presences and absences.

However, the main concern in using spatially non-representative data is that these models will likely lead to incorrect and biased predictions. The model performance statistics chosen may not reflect the true performance of the models in making predictions in cases where they are developed using differently distributed training and validation samples, as in our case. For the first time in any evaluation we have seen, our analysis quantifies the discrepancies of model performance statistics and model prediction success as influenced by the spatial pattern of sample points. (1) We can assess the effect of sampling bias on the scales selected in the multi-scale optimization. (2) We can compare the sets of covariates retained. (3) We can compare the signs of the coefficients. (4) We can compare the magnitude of the coefficients, between the empirical and the simulated models.

4.2. Bias in scale selection

The univariate analyses produced significant differences in the selection of the best scales between the spatially representative and the spatially non-representative simulations, as well as compared to the empirical model. Spatially non-representative data sets provided a more constant set of best scales, compared to the spatially representative simulations, where only one covariate showed the same best scale across the iterations. These results can be due to the more homogenous and clustered distribution of the spatially non-representative locations in the study area, compared to the random distribution of the spatially representative locations.

When compared to the best scales shown by the empirical model, the spatially non-representative simulations showed substantially greater accuracy than the spatially representative simulations. The covariates showing the same best scales across the spatially non-representative iterations were also coherent with the best scale of those covariates in the empirical model.

Overall, and contrary to our expectations, the non-representative simulated models produced predictions for each covariate that were more consistent across iterations and more similar to the scales of the empirical model, which was stipulated as the true probability surface for the simulations. We expected that sampling a broader range of ecological gradients in a representative way would improve the ability of the models to correctly identify the scales of relationship. However, sampling randomly also resulted in a higher proportion of samples falling in areas of low probability of presence, and the lower number of simulated detections likely resulted in lower ability to resolve the scale dependent relationships between each covariate and flammulated owl occurrence. These observations could have profound implications for citizen science, particularly common for avian species (e.g., eBird (Sullivan et al., 2009)). Data collected through citizen science projects have often been criticized for their spatial bias (Boakes et al., 2010). Although the risk of producing predictive maps that are biased towards easily accessible areas is still concrete (Kadmon et al., 2004), our results demonstrated that spatially non-representative data can provide accurate predictions of species-habitat relationships, encouraging the use of citizen science data collections.

4.3. Bias in covariates selected, signs and coefficients

The spatially representative and the spatially non-representative simulations showed discrepancies with the empirical model in the sets of covariates retained by the models, and in their coefficients. The spatially non-representative models again had much closer match to the empirical model on which they were trained. The non-representative models were more similar, on average, to the empirical model in terms of covariates selected, magnitude of coefficients and had fewer covariates that changed sign (i.e., direction of relationship).

Importantly, the non-representative models were more consistent

Table 5
Representative scales of the covariates from (a) the spatially representative and (b) the spatially non-representative models used for the cross-validations of the empirical model. Shown are the mean scales, the absolute differences between the mean scales and the empirical model's best scale and the *p*-value of the Wilcoxon test between the simulated models' best scales and the empirical model's best scale.

a)													
Covariate	**Best scale (m)**										**Mean (m)**	**Difference (m)**	**Wilcoxon test *p*-value**
DFc	5,000	5,000	5,000	5,000	4,900*	5,000	5,000	5,000	5,000	5,000	4,990	10	1.00E+00
DFed	5,000*	2,900*	4,700*	1,900*	4,500*	4,700*	5,000*	2,000*	4,400*	5,000*	4,010	1,410	4.24E-01
DFp	700*	800*	1,100*	700*	1,400*	700*	600*	1,600*	1,000*	800*	940	760	1.50E-01
DFpl	500*	500*	300*	400	1,300*	4,700*	400	500*	300*	2,600*	1,150	750	6.28E-01
PPc	1,700*	3,000*	2,700*	3,100	2,200*	2,900*	1,400*	1,000*	2,400*	2,700*	2,310	790	2.04E-01
PPed	4,600*	2,900*	200*	4,300*	5,000	4,300*	5,000	2,600*	400*	500*	2,980	2,020	2.63E-01
PPp	300*	300*	100	100	100	200*	200*	100	300*	300*	200	100	3.94E-01
PPpl	4,600*	200*	200*	5,000*	5,000*	3,600*	4,000*	4,600*	1,800*	700*	2,970	1,930	4.26E-01
Fc	5,000	4,300*	5,000	5,000	5,000	5,000	4,700*	5,000	5,000	5,000	4,900	100	8.15E-01
Fed	4,100*	2,800*	5,000	5,000	4,500*	5,000	5,000	5,000	3,700*	5,000	4,510	490	5.83E-01
Fp	2,200*	2,000*	900*	1,000*	2,100*	2,400*	1,500*	1,900*	1,000*	1,700*	1,670	3,230	1.54E-01
Fpl	200*	100*	100*	300*	600*	600*	400*	100*	200*	100*	270	4,730	1.43E-01
Han	400*	100*	1,800*	300*	700*	1,300*	300*	1,000*	200*	2,300*	840	4,160	1.54E-01
DEM	200*	400*	3,900*	600*	3,800*	300*	100	1,000*	900*	300*	1,150	1,050	2.04E-01
S	400*	1,300*	5,000*	800*	1,400*	5,000*	200*	3,800*	500*	2,600*	2,100	200	8.74E-01
A	900*	5,000*	400*	2,700*	3,000*	100*	800*	3,600*	3,900*	1,800*	2,220	1,220	9.09E-01
SP	1,200	1,200	900*	1,100*	1,200	1,200	1,200	1,200	1,200	1,200	1,160	40	8.15E-01
R	2,900*	2,200*	5,000*	5,000*	2,600*	2,100*	4,300	5,000*	1,200*	4,500*	3,480	820	1.00E+00
CDD	200*	5,000	2,100*	100*	100*	100*	300*	200*	100*	300*	850	4,150	1.92E-01
SpP	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	0	NA
SuP	3,500*	300*	200*	1,500*	5,000	5,000	2,000*	400*	2,200*	5,000	2,510	2,490	3.32E-01
b)													
Covariate	**Best scale (m)**										**Mean (m)**	**Difference (m)**	**Wilcoxon test *p*-value**
DFc	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	0	NA
DFed	5,000*	5,000*	5,000*	4,700*	5,000*	5,000*	200*	5,000*	4,900*	4,100*	4,390	1,790	2.27E-01
DFp	1,300*	1,400*	1,300*	1,700	900*	1,300*	1,600*	900*	1,600*	1,500*	1,350	350	1.99E-01
DFpl	600*	400	500*	500*	500*	500*	500*	500*	700*	600*	530	130	1.65E-01
PPc	1,900*	1,600*	1,600*	1,600*	1,800*	100*	3,100	2,100*	300*	1,800*	1,590	1,510	2.00E-01
PPed	200*	200*	300*	3,600*	100*	100*	100*	100*	5,000	100*	980	4,020	1.82E-01
PPp	100	100	100	100	100	100	100	100	100	100	100	0	NA
PPpl	100*	100*	100*	100*	100*	100*	100*	100*	5,000*	100*	590	4,310	1.01E-01
Fc	5,000	4,900*	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	4,990	10	1.00E+00
Fed	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	0	NA
Fp	4,700*	5,000*	5,000*	1,300*	5,000*	5,000*	5,000*	5,000*	5,000*	4,900	4,590	310	4.62E-01
Fpl	4,100*	4,700*	4,500*	3,900*	4,800*	5,000	5,000	5,000	5,000	5,000	4,700	300	4.90E-01
Han	5,000	4,800*	4,700*	3,800*	4,900*	5,000	5,000	5,000	5,000	5,000	4,820	180	5.83E-01
DEM	100	5,000*	5,000*	100	100	100	100	100	5,000*	100	1,570	1,470	6.83E-01
S	700*	1,800*	2,600*	5,000*	300*	600*	500*	2,200*	2,000*	400*	1,610	690	5.45E-01
A	2,300*	2,500*	2,000*	1,300*	1,900*	1,100*	1,000	3,500*	2,700*	1,300*	1,960	960	2.04E-01
SP	1,100*	1,200	1,300*	1,100*	1,200	1,200	1,200	1,200	1,200	1,200	1,190	10	1.00E+00
R	5,000*	3,400*	4,200*	5,000*	5,000*	3,300*	3,800*	3,000*	3,400*	5,000*	4,110	190	8.71E-01
CDD	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	0	NA
SpP	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	0	NA
SuP	5,000	5,000	5,000	4,500*	5,000	5,000	5,000	5,000	5,000	5,000	4,950	50	1.00E+00

* Scales different from the empirical model's best scales.

among iterations, and closer to the results of the empirical model in the covariates selected, and the magnitude and sign of coefficients. This was unexpected, as we thought that the spatially representative models would better capture the range of environmental conditions and provide more resolved predictions of the covariates driving the relationships and their relative effects (i.e., coefficients). We believe that the representative sampling actually was less able to resolve models due to the over-dispersion of the sampling locations, which covered many areas of low-quality habitat with very low probability of presence. The non-representative sampling might have performed better in producing more resolved and correct models because the sampling strategy was intentionally concentrated in areas with relatively high suitability for flammulated owl, and the training data had a better mixture of occurrences and absences, which improved the ability of the models to correctly fit the relationships inherent in the data.

4.4. Differences between model performance

Models built with spatially representative sampling greatly outperformed those built with non-representative sampling, based on standard measures such as AUC, PCC and Kappa statistics. However, models built with non-representative sampling more correctly identified the scales of relationship, the covariates in the models, and the magnitude and sign of the coefficients. This is a critical point. The apparent superiority of the representative models in model performance and the apparent superiority of the non-representative models in model prediction we believe are both caused by the same factors.

Specifically, the representative models had higher performance because the training and the validation datasets were both sampled over broad gradients with large differences between probability of presence in samples where flammulated owls were simulated to occur and those where they were simulated to not occur. This results in strong discrimination between presences and absences using AUC, PCC and Kappa statistics. At the same time, the non-representative models had sampling clustered in areas of the landscape with intermediate to high probability of presence, leading to lower ability to discriminate between presences and absences. As a result of the same sampling pattern, however, the non-representative models were more correct in terms of the scales, covariates and coefficients because, as noted above, they had a better balance between presences and absences clustered along ranges of the predictor covariates where the occurrence probabilities change along the threshold between presences and absences. This improves the model

Table 6
Model averaged parameter coefficients of (a) the spatially representative and (b) the spatially non-representative models used for the cross-validations of the empirical model. Shown are the average coefficients, the absolute differences between the average coefficients and the empirical model's coefficient and the *p*-value of the Wilcoxon test between the simulated models' coefficients and the empirical model's coefficient.

a)													
Covariate	**Coefficients**										**Mean**	**Difference**	**Wilcoxon test *p*-value**
Intercept	-3.52 E+00	-3.16 E+00	-3.65 E+00	-3.54 E+00	-3.43 E+00	-3.17 E+00	-3.24 E+00	-3.11 E+00	-3.29 E+00	-3.22 E+00	-3.33 E+00	0.08	1.00 E+00
DFc	1.25 E-02*	-5.34 E-02	-6.11 E-03	1.57 E-01*	8.97 E-04*	-3.12 E-01	-2.49 E-01	-3.43 E-02	-1.20 E-03	-1.69 E-01	-6.53 E-02	0.05	7.27 E-01
DFed	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA‡	NA
DFp	NA†	NA†	NA†	NA†	NA†	2.68 E-01	NA†	NA†	NA†	NA†	2.68 E-01	NA‡	NA
DFpl	6.94 E-01	5.73 E-01	6.81 E-01	7.85 E-01	7.07 E-01	NA†	4.53 E-01	6.83 E-01	5.11 E-01	3.98 E-01	6.09 E-01	0.07	8.00 E-01
PPc	3.04 E-01	3.90 E-01	4.91 E-01	6.39 E-01	3.88 E-01	5.44 E-01	3.93 E-01	1.93 E-01	3.49 E-01	4.44 E-01	4.14 E-01	0.06	7.27 E-01
PPed	-7.63 E-03*	1.04 E-03	1.36 E-01	3.62 E-01	2.02 E-01	2.25 E-01	2.98 E-03	1.46 E-01	1.71 E-01	NA†	1.38 E-01	0.10	8.00 E-01
PPp	2.02 E-01	2.22 E-01	1.85 E-01	1.22 E-01	2.05 E-01	3.37 E-02	6.39 E-03	1.09 E-01	NA†	1.98 E-01	1.43 E-01	0.04	1.00 E+00
PPpl	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA‡	NA
Fc	-2.66 E-01	-2.14 E-01	NA†	-2.28 E-02	-1.38 E-02	NA†	-2.30 E-01	-1.05 E-01	-1.42 E-02	-2.69 E-01	-1.42 E-01	0.06	1.00 E+00
Fed	NA†	NA†	2.71 E-01	NA†	NA†	3.74 E-02	NA†	NA†	NA†	NA†	1.54 E-01	NA‡	NA
Fp	NA†	2.99 E-02	2.77 E-03	1.34 E-02	NA†	NA†	1.52 E-03	-4.61 E-04	1.96 E-01	1.81 E-01	6.07 E-02	NA‡	NA
Fpl	NA†	NA†	2.03 E-02	NA†	NA†	3.07 E-01	NA†	9.11 E-03	NA†	8.51 E-02	1.05 E-01	NA‡	NA
Han	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA‡	NA
DEM	1.21 E-03*	NA†	-2.17 E-02	5.37 E-01*	-3.62 E-04	NA†	NA†	NA†	NA†	NA†	1.29 E-01	0.13	1.00 E+00
S	4.42 E-01	NA†	-9.32 E-03*	5.53 E-02	NA†	NA†	2.31 E-01	2.07 E-01	3.43 E-01	2.72 E-03	1.82 E-01	0.16	7.50 E-01
A	-3.04 E-01	-2.88 E-01	-3.31 E-01	-2.26 E-01	-1.12 E-01	-2.16 E-01	-2.18 E-01	-3.78 E-02	-2.40 E-01	-1.01 E-01	-2.07 E-01	0.04	7.27 E-01
SP	2.26 E-01	4.00 E-01	3.36 E-01	3.33 E-01	3.42 E-01	3.04 E-01	3.07 E-01	3.61 E-01	2.84 E-01	2.53 E-01	3.15 E-01	0.00	1.00 E+00
R	4.14 E-02	3.92 E-01	6.39 E-01	1.37 E-01	8.60 E-01	3.50 E-01	3.29 E-03	1.35 E-01	NA†	3.31 E-02	2.88 E-01	NA‡	NA
CDD	NA†	-3.01 E-03	NA†	NA†	NA†	-2.42 E-03	2.29 E-03	-6.46 E-03	-1.52 E-02	-1.00 E-01	-2.07 E-02	NA‡	NA
SpP	-2.83 E-01	1.98 E-03*	NA†	-4.46 E-01	-1.27 E-01	NA†	7.58 E-04*	NA†	-4.72 E-01	NA†	-2.21 E-01	0.02	1.00 E+00
SuP	-4.29 E-01	-3.39 E-01	-3.50 E-01	-5.14 E-01	-2.57 E-01	-7.98 E-02	-4.41 E-01	-2.31 E-01	-9.44 E-02	-5.00 E-01	-3.24 E-01	0.01	9.09 E-01
SAC	NA†	7.76 E+02	NA†	NA†	NA†	NA†	2.83 E+02	2.38 E+02	9.71 E+02	8.60 E+02	6.25 E+02	78.07	1.00 E+00
b)													
Covariate	**Coefficients**										**Mean**	**Difference**	**Wilcoxon test *p*-value**
Intercept	-1.93 E+00	-1.96 E+00	-1.96 E+00	-2.10 E+00	-1.97 E+00	-1.85 E+00	-2.02 E+00	-2.11 E+00	-2.00 E+00	-2.05 E+00	-1.99 E+00	1.26	1.82 E-01
DFc	-1.02 E-01	-2.10 E-02	2.52 E-04*	-2.91 E-01	-7.66 E-02	-8.87 E-03	7.13 E-03*	-4.95 E-03	5.49 E-02*	-4.79 E-02	-4.91 E-02	0.06	3.64 E-01
DFed	NA†	NA†	NA†	NA†	NA†	NA†	1.28 E-01	NA†	NA†	NA†	1.28 E-01	NA‡	NA
DFp	NA†	NA†	NA†	NA†	1.10 E-02	NA†	NA†	1.27 E-02	NA†	NA†	1.20 E-02	NA‡	NA
DFpl	5.24 E-01	4.18 E-01	5.59 E-01	4.98 E-01	4.99 E-01	5.76 E-01	4.30 E-01	6.58 E-01	6.30 E-01	5.94 E-01	5.39 E-01	0.00	1.00 E+00
PPc	1.85 E-01	2.09 E-01	2.20 E-01	2.14 E-01	2.33 E-01	2.41 E-02	2.62 E-01	1.62 E-01	1.18 E-03	1.86 E-01	1.70 E-01	0.18	1.82 E-01
PPed	1.85 E-01	1.46 E-02	2.17 E-01	2.39 E-04	NA†	NA†	NA†	NA†	8.00 E-02	NA†	9.94 E-02	0.07	1.00 E+00
PPp	1.40 E-01	2.78 E-01	1.90 E-01	2.17 E-01	2.05 E-01	2.63 E-01	2.25 E-01	2.56 E-01	2.37 E-01	3.21 E-01	2.33 E-01	0.05	3.64 E-01
PPpl	NA†	NA†	NA†	7.29 E-03	NA†	NA†	NA†	NA†	NA†	NA†	7.00 E-03	NA‡	NA
Fc	NA†	NA†	NA†	NA†	-2.47 E-01	-6.97 E-04	-1.07 E-01	NA†	-1.39 E-01	-3.05 E-01	-1.60 E-01	0.04	1.00 E+00
Fed	6.51 E-02	3.81 E-02	-1.12 E-03	1.04 E-01	NA†	NA†	NA†	2.07 E-01	NA†	NA†	8.26 E-02	NA‡	NA
Fp	-1.49 E-01	-1.70 E-03	4.99 E-03	1.18 E-02	NA†	NA†	NA†	-6.97 E-03	NA†	NA†	-2.82 E-02	NA‡	NA

(*continued on next page*)

Table 6 (*continued*)

Fpl	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA‡	NA
Han	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA†	NA‡	NA
DEM	NA†	3.29 E-03*	1.35 E-03*	NA†	-7.26 E-04	1.36 E-02*	9.42 E-02*	-4.44 E-04	-7.77 E-02	-3.59 E-03	3.63 E-03	0.00	1.00 E+00
S	8.73 E-02	2.65 E-01	NA†	NA†	1.06 E-01	1.51 E-01	2.29 E-01	2.37 E-01	1.86 E-01	3.78 E-02	1.62 E-01	0.18	2.22 E-01
A	-1.31 E-01	-1.22 E-01	-1.85 E-01	-2.02 E-01	-1.64 E-01	-1.71 E-01	-1.99 E-01	-1.55 E-02	-1.12 E-01	-1.06 E-01	-1.41 E-01	0.03	7.51 E-01
SP	2.68 E-01	4.33 E-01	3.91 E-01	3.89 E-01	4.27 E-01	2.91 E-01	3.88 E-01	3.07 E-01	4.31 E-01	3.84 E-01	3.71 E-01	0.06	7.27 E-01
R	NA†	NA†	1.46 E-01	2.28 E-01	5.04 E-02	NA†	NA†	NA†	NA†	5.61 E-02	1.20 E-01	NA‡	NA
CDD	2.19 E-03	NA†	NA†	-1.88 E-03	2.80 E-03	-9.22 E-03	9.90 E-02	1.21 E-03	NA†	4.17 E-02	1.94 E-02	NA‡	NA
SpP	-6.10 E-02	-2.07 E-01	-1.32 E-01	-6.11 E-02	-1.03 E-03	-1.64 E-01	-2.08 E-01	-1.70 E-03	-1.37 E-01	-1.76 E-03	-9.75 E-02	0.10	4.27 E-01
SuP	-1.49 E-01	-2.68 E-01	-2.47 E-01	-2.33 E-01	-2.28 E-01	-3.02 E-01	-2.17 E-01	-1.37 E-01	-2.29 E-01	-5.83 E-02	-2.07 E-01	0.10	1.82 E-01
SAC	NA†	NA†	NA†	NA†	NA†	NA†	9.39 E+00	NA†	5.67 E+01	2.42 E+01	3.01 E+01	673.41	5.00 E-01

† Covariate not evaluated because dropped after multicollinearity analysis.
‡ Difference not calculated because the covariate was dropped in the empirical model.
* Coefficient showing an opposite sign from the empirical model's coefficient.

Table 7
Models performances for the spatially representative and the spatially non-representative models. In order to maximize models' Kappa statistics, different optimal thresholds were used.

	Spatially representative						Spatially non-representative					
	Threshold	PCC	Kappa	Sensitivity	Specificity	AUC	Threshold	PCC	Kappa	Sensitivity	Specificity	AUC
Model 1	0.21	0.91	0.35	0.47	0.94	0.83	0.37	0.62	0.16	0.59	0.63	0.65
Model 2	0.24	0.92	0.34	0.40	0.96	0.82	0.43	0.73	0.18	0.34	0.83	0.66
Model 3	0.23	0.91	0.30	0.39	0.95	0.81	0.40	0.62	0.18	0.65	0.61	0.66
Model 4	0.19	0.90	0.32	0.47	0.93	0.83	0.48	0.71	0.14	0.33	0.81	0.63
Model 5	0.30	0.92	0.33	0.39	0.96	0.81	0.51	0.74	0.19	0.34	0.84	0.67
Model 6	0.23	0.91	0.25	0.29	0.95	0.79	0.50	0.74	0.18	0.32	0.85	0.66
Model 7	0.19	0.91	0.31	0.43	0.94	0.82	0.46	0.73	0.19	0.35	0.83	0.67
Model 8	0.26	0.92	0.34	0.39	0.96	0.83	0.52	0.75	0.20	0.32	0.87	0.67
Model 9	0.18	0.92	0.37	0.44	0.95	0.83	0.38	0.62	0.18	0.65	0.60	0.67
Model 10	0.17	0.89	0.29	0.50	0.91	0.81	0.49	0.73	0.19	0.35	0.84	0.67
Mean	0.22	0.91	0.32	0.42	0.95	0.82	0.45	0.70	0.18	0.42	0.77	0.66
(± SD)	(± 0.04)	(± 0.01)	(± 0.03)	(± 0.06)	(± 0.02)	(± 0.01)	(± 0.06)	(± 0.06)	(± 0.02)	(± 0.14)	(± 0.11)	(± 0.01)

algorithm's ability to correctly identify scales, covariates, magnitude and sign of the coefficients.

These results have several interesting and important implications. First, they suggest that spatially non-representative models may sometimes perform better than would be indicated by the model performance measures of AUC, PCC and Kappa statistics. In our analysis, these models were better than the spatially representative models in terms of their match to the empirical model on which they were trained. Hence, their apparent low performance is an artefact of the clustered nature of the training and validation samples, and is not due to the model's quality itself. Second, they suggest that the empirical model we produced and on which we based our simulations is likely also much better than would be indicated by AUC, PCC and Kappa statistics, as it was trained on the same spatial pattern of covariates as the non-representative models. Collectively, these results suggest that spatially representative sampling may not improve models built with presence-absence data in cases of over-dispersion of sampling locations, as may frequently happen when sampling for rare species with low extent of suitable habitat. In such cases, our results suggest that spatially non-representative sampling, clustered in ranges of environmental gradients across which presence-absence pattern pivots, may be most effective and efficient.

Another important insight from this exercise was that spatial scaling analysis and logistic regression were not able to correctly identify the scales, covariates or magnitudes of coefficients with nearly as much success as we expected. By using a simulation approach, we produced a pattern of occurrence probability with known relationships to covariates, used that probability to produce large representative presence-absence samples and then trained predictive models using a scale-optimized modelling framework. We expected that this modelling framework would show very high ability to extract the true scales, covariates and coefficients. The fact that neither the representative nor the non-representative models had very high precision in identifying these parameters is disconcerting, since this framework is routinely applied, and management and conservation decisions are widely based on these predictions. In both the representative and non-representative models there were frequent errors in identifying the correct scales of covariates (73% for representative and 53% for non-representative), the correct covariates (8.57% omitted and 13.81% committed for representative, and 6.67% omitted and 11.90% committed for non-representative), the magnitude of the coefficients (0.06 average difference for representative and 0.16 for non-representative) and their sign (6.38% mismatches in coefficients' sign for representative and 4.96% for non-representative). Further work using simulation approaches such as used here is warranted to more fully evaluate the ability of species distribution models to correctly identify scales, driving covariates, magnitudes and signs of relationships of species presence-absence patterns.

4.5. Ecological and conservation implications for flammulated owl

Based on the observation that the non-representative simulated datasets produced models that perform much better than indicated by standard model performance statistics, we believe that our empirical model more reliably reflects the habitat suitability for flammulated owls than may be indicated by its AUC, for example. The model closely matches past findings on the criticality of Douglas fir and ponderosa pine forests for the species. Scholer et al. (2014) found Douglas fir and ponderosa pine forests to be important predictors, respectively at fine scale (400m) and at mid-broad scale (3km), in line with our results on the best scales for the percent of landscape covered by Douglas fir and by ponderosa pine. In addition, percent of the landscape covered by Douglas fir was the most important covariate in the model, supporting previous studies on the importance of Douglas fir to the presence of the species, which likely provides the right combination of park-like stands and open forest needed by flammulated owls (Christie and van Woudenberg, 1997; McCallum, 1994a; Scholer et al., 2014). Flammulated owls occupy contiguous tracts of Douglas fir and ponderosa pine forests, or a combination of the two, surrounded by a matrix of homogenous cover types (McCallum, 1994b; Scholer et al., 2014). Our model supports these findings, demonstrating that the aggregation indexes for Douglas fir, ponderosa pine and forest have their strongest influence at broad scales. We also confirmed previous findings on the topographic features selected by flammulated owls, highlighting that aspect and roughness were selected at mid-broad scales, similarly to what Scholer et al. (2014) found. Negative selection for aspect indicates a selection for south-facing slopes, in agreement with observations made by Bull et al. (1990) and by Barnes (2007), while positive selection for slope position indicates a preference for ridgetops which, combined with the south-facing slopes, are associated with open canopy and park-like stands. In addition, warmer temperatures on south slopes are needed for physiological demands and earlier release of snow packs, creating favourable conditions for insect prey. All climatic covariates were selected at the broadest scale, revealing their large-scale effects on flammulated owl habitat selection. Only spring and summer precipitation were retained in the final model, and both showed a negative coefficient, demonstrating their negative effect on the species presence. Overall, we predicted that flammulated owl presence probability in the United States Northern Rocky Mountains is highest in Douglas fir and ponderosa pine forest, on south aspects and upper slopes, having relatively open canopy and high solar exposure.

5. Funding

This work was supported by the United States Forest Service Northern Region and by the United States Forest Service Rocky Mountain Research Station.

Credit Author Statement

S.A.C. conceived the study and designed the statistical analysis. B.H. and A.C. designed and supervised the fieldwork. T.N.W. led the fieldwork. L.C. and H.Y.W. prepared the data for analysis. L.C. analysed and interpreted data with substantial contribution of S.A.C. and H.Y.W. L.C. led the writing of the manuscript with substantial input of S.A.C. All the authors contributed critically to the drafts and gave final approval for publication.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ecolmodel.2021.109566.

References

Barnes, K.P., 2007. Ecology, Habitat Use, and Probability of Detection of Flammulated Owls in the Boise National Forest. Boise State University, Boise, ID, USA.
Bartoń, K., 2013. MuMIn: Multi-model inference.
Bates, D., Machler, M., Bolker, B.M., Walker, S.C., 2015. Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw 67, 1–48.
Bivand, R., Piras, G., 2015. Comparing Implementations of Estimation Methods for Spatial Econometrics. J Stat Softw 63, 1–36.
Boakes, E.H., McGowan, P.J.K., Fuller, R.A., Ding, C.Q., Clark, N.E., O'Connor, K., Mace, G.M., 2010. Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data. Plos Biol 8.
Bull, E.L., Wright, A.L., Henjum, M.G., 1990. Nesting Habitat of Flammulated Owls in Oregon. J Raptor Res 24, 52–55.
Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer, New York, NY, USA.
Christie, D.A., van Woudenberg, A.M., 1997. Modeling Critical Habitat for Flammulated Owls (Otus flammeolus). In: Duncan, J.R., Johnson, D.H., Nicholls, T.H. (Eds.), Modeling Critical Habitat for Flammulated Owls (Otus flammeolus). Biology and Conservation of Owls of the Northern Hemisphere. USDA Forest Service Gen. Tech. Rep. NC-190 97–106.
Cushman, S.A., Elliot, N.B., Macdonald, D.W., Loveridge, A.J., 2016. A multi-scale assessment of population connectivity in African lions (Panthera leo) in response to landscape change. Landscape Ecol 31, 1337–1353.
Cushman, S.A., Macdonald, E., Landguth, E., Malhi, Y., Macdonald, D., 2017. Multiple-scale prediction of forest loss risk across Borneo. Landscape Ecol 32, 1581–1598.
Evans, J.S., Oakleaf, J., Cushman, S.A., Theobald, D., 2014. An ArcGIS toolbox for surface gradient and geomorphometric modeling, version 2.0-0. Available: http://evansmurphy.wix.com/evansspatial.
Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ Conserv 24, 38–49.
Fourcade, Y., Engler, J.O., Rodder, D., Secondi, J., 2014. Mapping Species Distributions with MAXENT Using a Geographically Biased Sample of Presence Data: A Performance Assessment of Methods for Correcting Sampling Bias. Plos One 9.
Freeman, E., Moisen, G., 2007. PresenceAbsence: An R Package for Presence Absence Analysis. J Stat Softw 23.
Fylling, M.A., Carlisle, J.D., Cilimburg, A.B., Blakesley, J.A., Linkhart, B.D., Holt, D.W., 2010. Partners in Flight – Western Working Group. Flammulated owl survey protocol. http://sites.google.com/site/pifwesternworkinggroup/projects/flammulated-owl-monitoring.
Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol Model 135, 147–186.
Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G., 2013. High-Resolution Global Maps of 21st-Century Forest Cover Change. Science 342, 850–853.
Hirzel, A., Guisan, A., 2002. Which is the optimal sampling strategy for habitat suitability modelling. Ecol Model 157, 331–341.
Jenness, J., 2013. DEM Surface Tools for ArcGIS. Jenness Enterprises. Available at. http://www.jennessent.com/arcgis/surface_area.htm.
Johnson, R.R., Brown, B.T., Haight, L.T., Simpson, J.M., 1981. Playback Recordings as a Special Avian Censusing Technique. Studies in Avian Biology 6, 68–75.
Kadmon, R., Farber, O., Danin, A., 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecol Appl 14, 401–413.
Kramer-Schadt, S., Niedballa, J., Pilgrim, J.D., Schroder, B., Lindenborn, J., Reinfelder, V., Stillfried, M., Heckmann, I., Scharf, A.K., Augeri, D.M., Cheyne, S.M., Hearn, A.J., Ross, J., Macdonald, D.W., Mathai, J., Eaton, J., Marshall, A.J., Semiadi, G., Rustam, R., Bernard, H., Alfred, R., Samejima, H., Duckworth, J.W., Breitenmoser-Wuersten, C., Belant, J.L., Hofer, H., Wilting, A., 2013. The importance of correcting for sampling bias in MaxEnt species distribution models. Diversity and Distributions 19, 1366–1379.
LANDFIRE, 2005. Existing Vegetation Type Layer and Digital Elevation Model Layer. Department of the Interior, Geological Survey (Online), U.S. http://landfire.cr.usgs.gov/viewer/.
Levin, S.A., 1992. The Problem of Pattern and Scale in Ecology. Ecology 73, 1943–1967.
MacKenzie, D.I., 2005. What are the issues with presence-absence data for wildlife managers? J Wildlife Manage 69, 849–860.
Mateo-Sánchez, M.C., Cushman, S.A., Saura, S., 2014. Scale dependence in habitat selection: the case of the endangered brown bear (Ursus arctos) in the Cantabrian Range (NW Spain). Int J Geogr Inf Sci 28, 1531–1546.
McCallum, D.A., 1994a. Flammulated Owl (Otus flammeolus. In: Poole, A. (Ed.), The Birds of North America. Cornell Lab of Ornithology, Ithaca, NY, USA. http://bna.birds.cornell.edu/bna/species/093.
Review of Technical Knowledge: Flammulated Owls McCallum, D.A., 1994b. Flammulated, Boreal, and Great Grey Owls in the United States: A Technical Conservation Assessment. In: Hayward, G.D., Verner, J. (Eds.), Flammulated, Boreal, and Great Grey Owls in the United States: A Technical Conservation Assessment. USDA Forest Service Gen. Tech. Rep. RM-253, Rocky Mountain Forest and Range Experiment Station 14–46.

McGarigal, K., Cushman, S.A., Ene, E., 2012. FRAGSTATS v4: spatial pattern analysis program for categorical and continuous maps. Computer software program produced by the authors at the University of Massachusetts, Amherst, MA. Available: http://www.umass.edu/landeco/research/fragstats/fragstats.html.

McGarigal, K., Wan, H.Y., Zeller, K.A., Timm, B.C., Cushman, S.A., 2016. Multi-scale habitat selection modeling: a review and outlook. Landscape Ecol 31, 1161–1175.

Merow, C., Smith, M.J., Silander, J.A., 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography 36, 1058–1069.

Mohler, C.L., 1983. Effect of Sampling Pattern on Estimation of Species Distributions Along Gradients. Vegetatio 54, 97–102.

Phillips, S.J., Dudik, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol Appl 19, 181–197.

Pliscoff, P., Luebert, F., Hilger, H.H., Guisan, A., 2014. Effects of alternative sets of climatic predictors on species distribution models and associated estimates of extinction risk: A test with plants in an arid environment. Ecol Model 288, 166–177.

PRISM Climate Group, 2016. Oregon State University. http://prism.oregonstate.edu.

Core Team, R, 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL. https://www.R-project.org/.

Rathbun, S.L., Gerritsen, J., 2001. Statistical Issues for Sampling Wetlands. In: Rader, R. B., Batzer, D.P., Wissinger, S.A. (Eds.), Bioassessment and Management of North American Freshwater Wetlands. Wiley, New York, NY, USA, pp. 45–58.

Scholer, M.N., Leu, M., Belthoff, J.R., 2014. Factors Associated with Flammulated Owl and Northern Saw-Whet Owl Occupancy in Southern Idaho. J Raptor Res 48, 128–141.

Shirk, A.J., Raphael, M.G., Cushman, S.A., 2014. Spatiotemporal variation in resource selection: insights from the American marten (Martes americana). Ecol Appl 24, 1434–1444.

Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D., Kelling, S., 2009. eBird: A citizen-based bird observation network in the biological sciences. Biol Conserv 142, 2282–2292.

van Proosdij, A.S.J., Sosef, M.S.M., Wieringa, J.J., Raes, N., 2016. Minimum required number of specimen records to develop accurate species distribution models. Ecography 39, 542–552.

Varela, S., Anderson, R.P., Garcia-Valdes, R., Fernandez-Gonzalez, F., 2014. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. Ecography 37, 1084–1091.

Vergara, M., Cushman, S.A., Urra, F., Ruiz-Gonzalez, A., 2016. Shaken but not stirred: multiscale habitat suitability modeling of sympatric marten species (Martes martes and Martes foina) in the northern Iberian Peninsula. Landscape Ecol 31, 1241–1260.

Wan, H.Y., McGarigal, K., Ganey, J.L., Lauret, V., Timm, B.C., Cushman, S.A., 2017. Meta-replication reveals nonstationarity in multi-scale habitat selection of Mexican Spotted Owl. Condor 119, 641–658.

Wessels, K.J., Van Jaarsveld, A.S., Grimbeek, J.D., Van der Linde, M.J., 1998. An evaluation of the gradsect biological survey method. Biodivers Conserv 7, 1093–1121.

Wiens, J.A., 1989. Spatial Scaling in Ecology. Funct Ecol 3, 385–397.

Williams, K.J., Belbin, L., Austin, M.P., Stein, J.L., Ferrier, S., 2012. Which environmental variables should I use in my biodiversity model? Int J Geogr Inf Sci 26, 2009–2047.

Wisdom, M.J., Holthausen, R.S., Wales, B.C., Hargis, C.D., Saab, V.A., Lee, D.C., Hann, W.J., Rich, T.D., Rowland, M.M., Murphy, W.J., Eames, M.R., 2000. Source Habitats for Terrestrial Vertebrates of Focus in the Interior Columbia Basin: Broad-Scale Trends and Management Implications. USDA Forest Service Gen. Tech. Rep. GTR-485. Pacific Northwest Research Station, Portland, OR, USA.

Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., Smith, G.M., 2009. Mixed Effect Models and Extensions in Ecology with R. Springer, New York, NY, USA.