Can singing rate be used to predict male breeding status of forest songbirds? A comparison of three calibration models

EMILY J. UPHAM-MILLS,[1],† JODY R. REIMER [iD],[1,2] SAMUEL HACHÉ,[1,3] SUBHASH R. LELE,[2] AND ERIN M. BAYNE[1]

[1]Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E9 Canada
[2]Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta T6G 2G1 Canada
[3]Environment & Climate Change Canada, Yellowknife, Northwest Territories X1A 2P7 Canada

Abstract. For male songbirds, song rate varies throughout the breeding season and is correlated with breeding cycle stages. Although these patterns have been well documented, this relationship has not been used to predict a bird's breeding status from acoustic monitoring. This challenge of using a response (i.e., behavior) to indirectly measure an underlying biological state is common in ecology, but correctly addressing the associated statistical challenge of calibration is rare. The objective of this study was to determine whether variation in song rate can be used to predict the breeding status of the Olive-sided Flycatcher (*Contopus cooperi*). In 2016, song rates from 28 male Olive-sided Flycatchers were collected from human observers ($n = 545$ five-minute counts) and breeding status (i.e., single, paired, and feeding young) was monitored throughout the breeding season. The predictive ability of three modeling approaches—regression, hierarchical, and a classification tree—was evaluated using sensitivity and specificity to determine the best modeling approach. The hierarchical model was the best at predicting all three breeding status classes, with a mean sensitivity of 69%, compared with 54% and 50% from the regression and machine learning models, respectively. Our results suggest that song rate can be used as an indirect measurement of breeding status in the Olive-sided Flycatcher when using a hierarchical modeling approach to calibrate the breeding status–song rate relationship. This novel modeling approach provides a cost-effective tool to collect much needed demographic information over large spatial extents and inform species status assessments, recovery strategies, and management plans for species of conservation concern.

† **E-mail:** uphammil@ualberta.ca

INTRODUCTION

Ecologists often desire information on the state of an organism or environment that can be challenging to measure directly. Because of this, many state variables require indirect measurements (Stephens et al. 2015). For example, leaf area index, an important metric of forest function, is often measured indirectly through recording light absorption patterns because of the high cost of directly measuring leaf dimensions (Olivas et al. 2013). Other examples include using indicator species to track changes in the state of the environment (Lindenmayer and Likens 2011) or satellite tracking data as an indirect measure of wildlife feeding behavior (Robinson et al. 2007). Although the use of indirect metrics is common in ecology (Stephens et al. 2015), the statistical methods used to infer the relationship between the state variable of interest and the

indirect proxy are often oversimplified. For simplicity, it may be tempting to ignore causal dependencies when analyzing these relationships. However, this might result in incorrect conclusions or low predictive accuracy due to the error distributions implied by standard statistical models (e.g., regression models). Statistical calibration aims to estimate an independent variable (the cause) from a dependent variable (the effect; Osborne 1991). Despite acknowledgment of the importance of calibration in some fields (e.g., water quality health; ter Braak and Barendregt 1986, Hall and Smol 1992; and paleoecology, ter Braak 1995), relatively few ecological studies have used such approaches for creating effective indirect measurement techniques (Biondi and Waikul 2004).

In avian ecology, calibration models describing the relationship between a male songbird's breeding status and his behavior (cause and effect, respectively) may provide a novel way to monitor male breeding status indirectly. This information is required to inform sound conservation planning (Anders and Marshall 2005), but direct measurements of pairing success, nest success, and fledging rates for songbirds are expensive and logistically challenging to obtain (Martin et al. 1997). Thus, empirical data on breeding status are only available for a few species and over relatively small spatial extents (Holmes et al. 1992, Christoferson and Morrison 2001, Dussourd and Ritchison 2003, Haché et al. 2013). Indirect measures of breeding status, such as observing non-agonistic behavior toward conspecifics to confirm pairing status and observing adult birds carrying food to confirm the presence of young, have been suggested as an approach to decrease time and effort to estimate metrics such as fledging success (Vickery et al. 1992, Hunt et al. 2017). However, such methods are still time-consuming for many species occurring at low abundance and with large breeding territories and have not yet been rigorously calibrated.

We propose a simpler indirect measure of a songbird's breeding status: inferring breeding status from singing behavior. Songs in passerines are primarily used by a male to attract females and to defend a territory against conspecifics (Thorpe 1961, Armstrong 1973). For many species, males tend to sing at high rates when they are unpaired, with declines in singing rate as

their breeding status changes (i.e., unpaired to paired, mated to nest building, egg laying to incubating, incubating to feeding nestlings, etc.; Gibbs and Wenny 1993, Dussourd and Ritchison 2003, Liu et al. 2007). We refer to this pattern as the breeding status–song rate relationship or the "BSSR" relationship. While several studies have described the BSSR relationship, none have attempted to use song rate to predict breeding status (although Staicer et al. 2006, suggested the possibility). We explored three different calibration model types to assess the use of song rate to predict breeding status (single, paired, and feeding young) of the Olive-sided Flycatcher (*Contopus cooperi*; OSFL). Specifically, our objective was to test predictive ability of these three BSSR calibration models. We used OSFL song rate data from a study conducted in the Northwest Territories and northern Alberta, Canada. This species is designated as Threatened under Canada's Species at Risk Act (SARA 1994) and has experienced an overall population decline of 70% between 1970 and 2015 (Environment and Climate Change Canada 2017). Therefore, finding a cost-effective way to monitor breeding success is a priority in recovery planning (Environment and Climate Change Canada 2017).

Previous work has suggested that song rates and detection probabilities for the OSFL are influenced by breeding status, time of day, and day of year (Wright 1997). Time of day is an important predictor for singing activity in songbirds, as males sing the most around sunrise and song production declines throughout the day (Stacier et al. 1996). Date is also an important variable for predicting breeding status, because migratory birds will be single upon first arriving at the breeding ground, then will be more likely to be paired or have active nests as the days advance. Furthermore, latitude may affect breeding timing because more northern breeding sites will have later arrival times. We therefore tested the importance of each of these predictors in the BSSR models.

First, we performed a multinomial logistic regression of breeding status against song rate and temporal covariates. Arguably the simplest model we considered, this model conflates the causal relationship assumed between breeding status and song rate, and the causal mechanisms behind the temporal predictors. Secondly, we

used a hierarchical model, defined as a sequence of probability models arranged to describe conditionally dependent random variables (Kéry and Royle 2016). This modeling approach is useful for complex ecological modeling because of its ability to account for multiple sources of uncertainty (Cressie et al. 2009). In the case of BSSR calibration, a hierarchical model can be used to deconstruct the cause-and-effect relationships into one component that accounts for temporal variation in breeding status probabilities throughout the breeding season and a second component that models how breeding status and time of day affect song rate. Our third approach was to use a classification and regression tree (CART) model (Brieman et al. 1984), to predict breeding status from song rate, time of day, date, and latitude. Classification and regression tree is a machine learning approach, which has been recommended as a powerful method for modeling complex ecological data because of its ability to deal with nonlinear relationships and high-order interactions (De'Ath and Fabricius 2000). Classification and regression tree models are comprised of a series of binary splits, based on predictor variable values, to partition data into smaller groups and increase the proportion of any one class (i.e., categorical value) in each group (Kuhn and Johnson 2016).

Our objective was to determine the best modeling framework to accurately predict a songbird's breeding status. We measured the relative success of the three models by comparing prediction sensitivity and specificity for the three breeding status classes using K-fold cross-validation (Arlot and Celisse 2010). We conclude by discussing the strengths and weaknesses of the top-performing model and how this model may be further developed for use with autonomously recorded acoustic data.

METHODS

Study species

The OSFL is a neotropical migratory songbird that typically breeds in lowland coniferous forests of the Canadian boreal and montane coniferous or coastal regions of the Pacific Coast of Canada and the United States (Altman and Sallabanks 2012). In boreal regions, this aerial insectivore defends territories in conifer stands, recently burned forest, and shrubby patches (Haché et al. 2014). This species is often associated with edge habitat, such as forest edges beside wetlands or riparian areas, and man-made clearings (Altman and Sallabanks 2012). Olive-sided Flycatchers are generally monogamous birds that defend large territories (up to 40–45 ha; Altman and Sallabanks 2012). Extra-pair copulation rates are expected to be very low (Altman and Sallabanks 2012), as evidence of potential polygyny has only ever been recorded once (B. Altman, *unpublished manuscript*), and territories are large and typically located at least >100 m apart (Robertson et al. 2009). Adults usually build an open-cup nest in branches near the top of tall live or dead conifer trees, lay 3–4 eggs, and only lay one clutch per year (Altman and Sallabanks 2012). In northern latitudes (i.e., Alaska and central Alberta), spring arrival dates are mid- to late May (Salt and Salt 1976, Kessel and Gibson 1978, Wright 1997), and in Alaska, clutches initiate late May to mid-June with fledging in mid-July (Wright 1997).

Olive-sided Flycatchers sing a single song type, described as a loud, clear whistle or the mnemonic "quick, three-beers!" (Peterson 1980; Fig. 1). Although some variation exists among individuals (Robertson et al. 2009), and OSFLs may produce other call types for purposes such as alarming or courtship (i.e., "pips," "churrs," and "purrs"; Wright 1997), in this study we define a single song as the full, three-syllable "quick, three-beers."

A detailed study of OSFL singing behavior in Alaska revealed seasonal and daily patterns
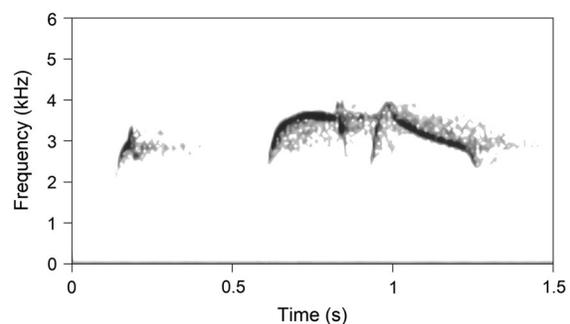

Fig. 1. Spectrogram of an Olive-sided Flycatcher song, with frequency in kilohertz (kHz) and time in seconds (s).

(Wright 1997). Seasonally, males sang regularly until they attracted a female, after which singing decreased dramatically until incubation. After hatching, males sang at an extremely low rate; however, males that remained unpaired sang regularly throughout the season. On a daily scale, males sang at very high rates of >8 songs/min early in the morning decreasing to 2 songs/min by mid-morning and between 1 and 2 songs/min from late morning to afternoon.

Study area

The study took place in northern Alberta and the Northwest Territories, Canada (Fig. 2) between 30 May and 22 July 2016. The sampling locations in Alberta were ~80 km north of Fort McMurray, in the Mid-Boreal Mixedwood Ecoregion (Strong and Leggat 1992) in stands comprised of upland jack pine (*Pinus banksiana)* forest, bog, and fen wetlands (dominated by black spruce, *Picea mariana* and tamarack, *Larix laricina*). Dominant understory shrubs included rose (*Rosa acicularis*), alder (*Alus* spp.), and aspen (*Populus tremuloides*) in upland sites and blueberry (*Vaccinium myrtilloides*), Labrador tea (*Ledum groenlandicum*), dwarf/bog birch (*Betula* spp.), and willow (*Salix* spp.) in lowland sites. The Northwest Territories study area ranged from ~30 km south of Fort Providence to Behchokò, with site access off Highway 3 (Fig. 2). This study area lies within the Great Slave Lowland Mid-Boreal Ecoregion (Ecosystem Classification Group 2007), which is dominated by wetlands (bogs and fens) and scattered patches of upland mixed-wood and jack pine forests. Some sampling locations were in stands that had burned in 2014 and 2015. Shrub species composition was similar to Alberta.

Sampling design

Sampling locations and time of sampling were selected based on known occurrence and behavioral observations of territorial male OSFLs during the previous two breeding seasons (Pardieck et al. 2016, Pankratz et al. 2017; E. Bayne, *unpublished data*; M. Knaggs, *unpublished data*). The spatial extent of our study area was selected to represent southern (latitude 57°) and northern locations (latitude 62°) to account for variation in daily activity levels. Between the last week of May and the first week of June 2016, potential

sampling locations were monitored using call playback surveys to confirm arrival and settlement of OSFLs. Call playback surveys consisted of 5 min of listening, followed by 30 s of playback, 2 min of listening, 30 s of playback, and a final 5 min of listening. If call playback surveys were conducted at a location on three separate days before 8 June 2016, and no OSFL was detected, or only once prior to the last visit, a location was deemed to not overlap a territory. Alternatively, if an OSFL was detected twice or more, the potential territory was considered occupied and included in our sampling locations. While monitoring potential territories, we detected additional males in other nearby locations (*n* = 5) and these locations were added to the sampling design. This resulted in 19 sampling locations in the Northwest Territories and 9 in Alberta.

For each territory, we conducted repeated visits approximately once per week, resulting in between 1 and 10 visits per bird (6 ± 2.43; mean ± standard deviation [SD]). During each visit, we assessed breeding status (i.e., single, paired [including being paired with no known nest, nest building, or incubating], or feeding young; Table 1). When breeding status could not be confirmed in the field, the status was back-calculated based on average breeding timing for the species (15 d for incubation and 19 d with nestlings; Wright 1997) using breeding status information from previous and subsequent visits. Dates when status could not be back-calculated with confidence (i.e., there were not enough dates with field-confirmed breeding status) were excluded from the analysis. A visit lasted 1 h and was conducted between sunrise and the first 6 h after sunrise. During each visit, we also measured song rate, defined as the mean singing rate based off four 5-min song counts. At time 0, 15, 30, and 45 min from the start of the visit, the number of songs produced by the OSFL was counted in a 5-min period. Song counts were only conducted when observers were close enough to see the male and were canceled if he flew away during the count period. Song count data were recorded until 30 June and 8 July, in Alberta and the Northwest Territories, respectively. In both study areas, territories were revisited once or twice between 8 and 22 July to confirm nest contents if not yet confirmed. Nests
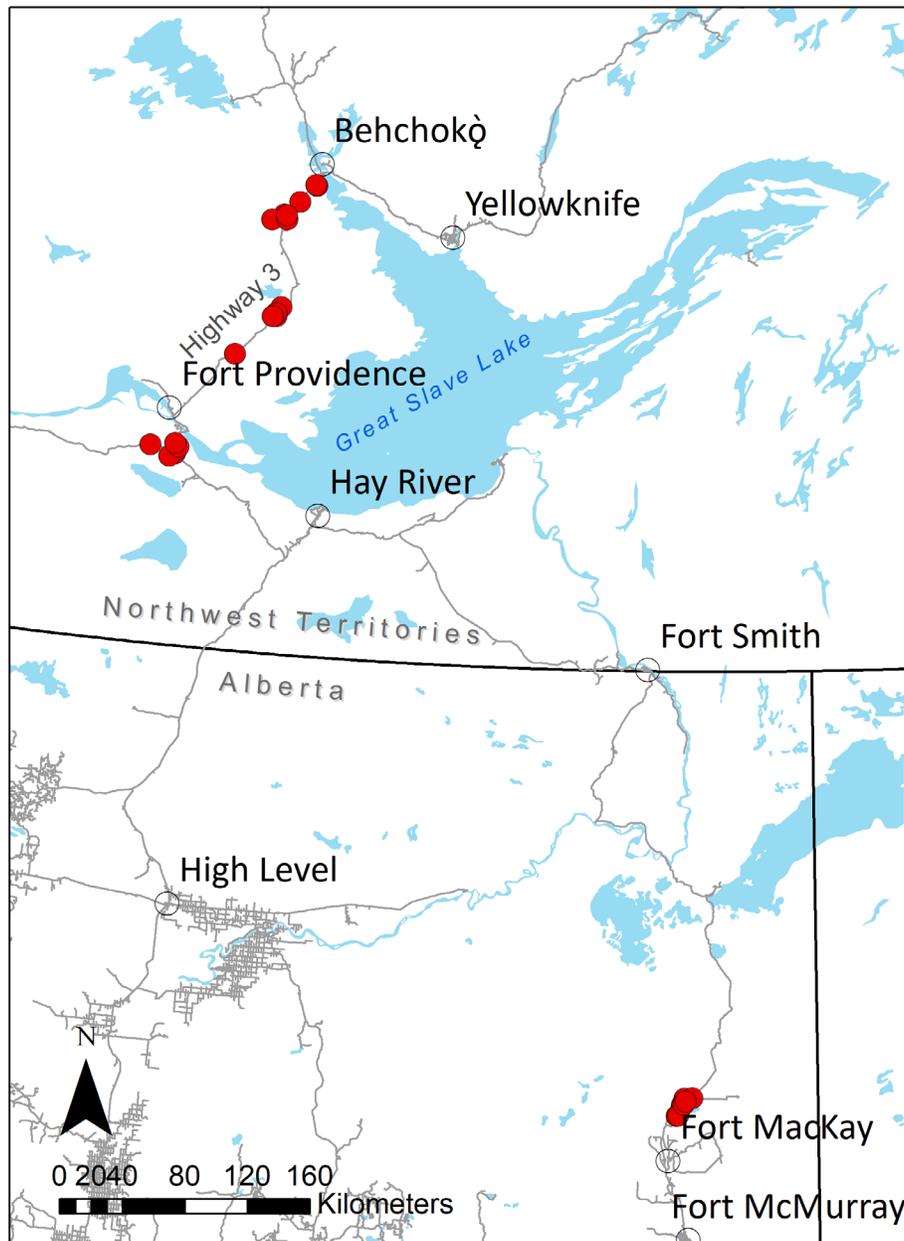
Fig. 2. Locations (red dots) where Olive-sided Flycatcher territories were monitored in 2016 in the Northwest Territories (n = 19) and northern Alberta (n = 9), Canada.

were located for 12 of the 28 males monitored in this study. For these males, nest contents were confirmed using a telescopic PVC pole (maximum height 17 m) with a video camera, which provided a live feed to a handheld monitor on the ground (The Peeper Cam, http://www.ibwo.org/camera.php, David Luneau, Arkansas, USA).

Methods of calibration

For the three model types, that is, multinomial logistic regression, hierarchical model, and classification tree, model selection was conducted on a set of candidate models allowing for identification of meaningful predictor variables. Song rate, time, ordinal date, and latitude were considered

Table 1. Breeding status classification descriptions and observational cues.

Classification	Description	Behavioral evidence
Single	Unpaired male defending a breeding territory	No interactions with a female detected on that day and during previous visits
Paired	Immediately after a female settled on a male's territory, no known nest	First observation of a female present on a male OSFL's territory, but nest not located
	Nest building: one or both adults seen building a nest	Either or both adults seen carrying nesting material
	Incubating: 15-d period pre-hatching when female spends most of her time sitting on eggs in the nest (Altman and Sallabanks 2012)	Female directly observed sitting on nest, male observed feeding a female on the nest or when nestlings were observed during a later visit (backdated incubation period based on hatch date)
Feeding young	One or more eggs have hatched, and nestlings are present	A parent observed holding/carrying insect and/or feeding an insect to nestlings, or direct observations of nestlings

Note: OSFL, Olive-sided Flycatcher.

as candidate predictors for each model type. Song rate was the average 5-min song count collected within one sampling hour ($n = 4$) during a visit. Time was calculated by subtracting time of sunrise from the mean time of the song counts. Both date and time were mean-centered and scaled by their SD. Latitude was a binary categorical variable, representing either the northern (Northwest Territories) or southern (Alberta) sites.

All models were built using the R statistical programming language (R Core Team 2017).

Multinomial logistic regression.—Multinomial logistic regression models were built using the multinom function in R package nnet (Venables and Ripley 2002, Ripley and Venables 2016). We compared six a priori models (Appendix S1: Table S1), ranging from the simplest models with a single predictor to the full model that included song rate, date, time as an interaction with song rate, and latitude as an interaction with date, as predictors of breeding status. We used the lowest Akaike information criterion (AIC) value (with a difference >2) to select the best-supported model (Burnham and Anderson 2002).

Hierarchical model.—The hierarchical modeling approach allowed us to relate the probability of an individual having a given breeding status, conditional on its song rate, to the probability of that individual having a given song rate, conditional on its breeding status. A popular approach to hierarchical model composition is to break down the process into multiple stages or component models, which make up the full model (Berliner 1996, Wikle 2003). We assembled a hierarchical model consisting of two components. The first component represented how temporal covariates (date and latitude, which may affect timing of breeding) influence the marginal probability of observing a bird in each breeding status:

$$\text{Component A.} \quad P(B|D, L)$$

where B is breeding status, D is date, and L is latitude. The second component represented the conditional probability of a singing rate given the breeding status and time of day:

$$\text{Component B.} \quad P(S|B, T)$$

where S is singing rate, and T is time. We assembled the two components to create a hierarchical model structured to determine the probability of each breeding status given the song rate, date, latitude, and time of day.

$$P(B|S, D, L, T) \propto \underbrace{P(B|D, L)}_{\text{Component A}} \underbrace{P(S|B, T)}_{\text{Component B}} \quad (1)$$

where $\propto$ means proportional to, with the constraint that the sum of the left-hand side probabilities for each breeding status must sum to 1 (see Appendix S2 for details regarding the Eq. 1 derivation). Eq. 1 describes all covariate relationships that we considered, but final models did not necessarily include all covariates. The breeding status variable in our model is known and observable, unlike state-space models (Patterson et al. 2008); hence, we were able to conduct

model selection on each component separately. This avoided the difficulty in model selection common with state-space models (i.e., where AIC is not properly defined and deviance information criterion, typically used with the Bayesian model selection, is problematic; Hooten and Hobbs 2015).

We first conducted model selection for component A, using the lowest AIC value within 2 to select the top model. For this component, we compared three a priori multinomial logistic regression (MLR) models relating the marginal probability of each breeding status to date and latitude (Appendix S1: Table S2).

For component B, we used a generalized linear model (GLM) for song rate, with time and breeding status as predictor variables. We considered 12 a priori models, using either breeding status as a single covariate or both breeding status and time. For both options, we tested a Poisson, a zero-inflated Poisson, a negative binomial, and a zero-inflated negative binomial song rate distribution (Appendix S1: Table S3). For component B, model sensitivity for the three predicted classes was used to select the top model, instead of AIC, with the purpose of maximizing model ability to predict individual breeding statuses from song rate.

Normal approximations of the parameter estimates and their standard error values from the top-ranked MLR (component A) and GLM (component B) were used as priors for the hyperparameters of model components. A common practice is to use so-called non-informative priors, but there are known issues with whether truly non-informative priors exist (Northrup and Gerber 2018; S. R. Lele, *unpublished manuscript*). We chose to use informative priors in an empirical Bayes framework as is suggested by Hamilton (1986) and Harris (1989) in the context of prediction.

Cumulatively, selected models for components A and B comprised the top hierarchical model used to predict breeding status class probability densities. These calibration distributions (i.e., posterior distributions) were generated using the Markov chain Monte Carlo methods from the package rjags in R (Plummer et al. 2016). We generated five Markov chains, discarding the first 1000 values as the burn-in, followed by 10,000 iterations. We used the Gelman-Rubin

diagnostic to test for convergence of the chains to a posterior distribution (Spiegelhalter et al. 1995, Brooks and Gelman 1998).

Classification tree.—The classification tree model (Brieman et al. 1984) was built using R package rpart (Therneau et al. 2018), using the Gini index as the impurity index (Wu et al. 2008). A set of classification trees were built to include a range of sizes, from unpruned (i.e., the tree with the highest number of branches, created using the default complexity parameter value of 0.01) to fully pruned (i.e., the tree with the smallest number of branches). We conducted model selection by choosing the classification tree that predicted with the highest mean sensitivity after K-fold cross-validation (i.e., leave-one-group-out, process described in more detail below). This model selection process is the best when the research objective behind the generation of classification trees is prediction (De'Ath and Fabricius 2000).

Model evaluation

Breeding status predictions were made using leave-one-group-out K-fold cross-validation, using one individual OSFL as the group-out. Specifically, all observations from one individual OSFL were removed from the dataset, leaving a training set with observations from 27 males, which was used to obtain predictive distributions. The model selection process was not repeated for each validation fold. Breeding status predictions were then made based on observations from the remaining 27 OSFL, and the process was repeated for each individual OSFL, resulting in a 28-fold cross-validation for each model type. The output from each model was a probability mass function for each sampling time, describing the probability of the individual having each of the three breeding statuses. We then used the breeding status with the highest probability as the predicted status.

We used the following performance statistics to compare the predictive ability from the top model in each model type: sensitivity (i.e., true-positive rate for each breeding status class), mean sensitivity (mean taken across the three breeding status classes), specificity (i.e., false-positive rate for each breeding status class), and mean specificity (mean taken across all three breeding status classes; see Appendix S1:

Table S4 for equations describing each prediction evaluation term). We used specificity and sensitivity as predictive measures because both are prevalence-independent test characteristics, meaning that their values do not depend on the prevalence of a value in the dataset. We tested whether the predicted breeding status prevalence was significantly different from the prevalence of observed breeding statuses by performing two tests of marginal homogeneity: a Bhapkar test for overall results (Bhapkar 1966) and the McNemar tests for each of the three predicted classes (McNemar 1947).

Sample prevalence

The dataset used in this study had an unequal prevalence of each breeding status to train the models; that is, ~70% of song rates were associated with paired observations and ~15% for both single and feeding young. This imbalance in the dataset occurred because the period between male arrival and pairing is brief relative to the amount of time required for nest building and incubating. Also, only a few song counts ($n = 23$) were collected from males in the feeding young stage because only 46% of the nests reached that status during our sampling period.

The challenge of unequal class representation is common in classification models, where predicting the class with the lowest representation is often of higher importance and interest (Ali et al. 2015). However, model algorithms often maximize accuracy (Ali et al. 2015) and are thus biased toward predicting the most prevalent class. A useful BSSR model should not be biased toward the paired status because the different breeding statuses provide important demographic information. Thus, we used the prevalence-independent metrics sensitivity and specificity for the three breeding statuses, individually and combined (i.e., mean sensitivity), to distinguish predictive performance for the different models.

RESULTS

We collected 20 ± 7.7 (mean $\pm$ SD; $n = 545$) five-minute song counts per territorial male, from which we calculated 160 mean 5-min song rates. The breeding statuses associated with these song rates were comprised of 26 (16.3%)

instances where males were classified as single, 111 (69.4%) as paired, and 23 (14.4%) as feeding young. The mean date of first detecting breeding males was 6 June (± 4.5 d; $n = 28$). Mean pairing date was 9 June (± 6.6; $n = 21$), start of incubation was 18 June (± 7.3 d; $n = 20$), and hatching was 2 July (SD = ± 7.6 d; $n = 17$).

Multinomial logistic regression

The top multinomial model included song rate and date as independent variables (Appendix S1: Table S1). Song rate had a significant negative effect on the probability of being paired vs. single (log odds ratio = 0.94, $P < 0.001$). Date had a significant positive effect on the probability of being paired (log odds ratio = 1.07, $P = 0.045$) or feeding young (log odds ratio = 1.74, $P < 0.001$) vs. single. OSFLs are more likely to be paired or feeding young than single later in the breeding season.

Hierarchical model

The top-ranked full model included time, date, and song rate (see Appendix S3 for BUGS-language script file for this model). For component A, breeding status was best modeled by date (Appendix S1: Table S2), where the probability of (1) being single is highest for early dates, (2) being paired is highest in the middle of the date range, and (3) feeding young is highest for later dates. The component B model that resulted in the highest mean sensitivity included both breeding status and time as predictors of song rate, using a Poisson distribution of mean-rounded song rate (see Appendix S1: Table S3 for contrasting AIC values).

Classification tree

The top classification tree model from the model selection (i.e., the one that best predicted all three breeding statuses) had four splits and included all predictor variables (ranked importance: date, song rate, time and latitude; Appendix S1: Fig. S1).

Model performance comparison

Based on the highest mean sensitivity, the top model type was the hierarchical model (69%), followed by the multinomial logistic regression and classification tree with mean sensitivity values of 54% and 50%, respectively (Table 2).

Table 2. Performance evaluation metrics of the three model types.

	Sensitivity				Specificity			
Model	Single	Paired	Feeding young	Mean	Single	Paired	Feeding young	Mean
MR	0.19	**0.92**	0.52	0.54	**0.96**	0.37	0.97	0.77
HM	**0.69**	0.50	**0.87**	**0.69**	0.80	**0.82**	0.78	**0.80**
CT	0.08	0.89	0.52	0.5	0.93	0.29	**0.98**	0.73

Notes: MR, multinomial logistic regression; HM, hierarchical model; CT, classification tree. The highest values for each metric and breeding status are highlighted in bold.

All three models overpredicted some breeding statuses, indicated by their specificity values (Table 2). The multinomial logistic regression and classification tree both overpredicted paired at a high rate (specificity values of 0.37 and 0.29, respectively) compared to the hierarchical model (0.82 specificity). The marginal frequencies (i.e., predicted breeding status prevalence) of the multinomial logistic model and the classification tree were similar (<10% predictions of single and feeding young and >80% for paired), while those of the hierarchical model predicted a lower prevalence of paired individuals (Fig. 3), with the true prevalence lying between these two extremes. Test results for marginal homogeneity between the predicted and the observed breeding statuses were similar for all three models, with the prevalence of single and paired differing significantly from true prevalence (Appendix S1: Table S4). The prevalence of feeding young did not differ significantly between predicted and true breeding statuses for the classification tree and multinomial logistic regression, but the predicted value differed significantly from the true breeding status for the hierarchical model.

Discussion

The top modeling approach in our comparison was the hierarchical model (hereafter referred to as the BSSR model), which predicted all three breeding statuses correctly at a higher rate (i.e., mean sensitivity) than the regression and CART models and was less prone to overpredict any given breeding status (i.e., mean specificity). The challenge with measuring breeding status indirectly is the statistical calibration of the underlying behavioral mechanism when breeding status causes changes in singing rate. The hierarchical structure of the BSSR model allowed us to address this challenge while also accounting for
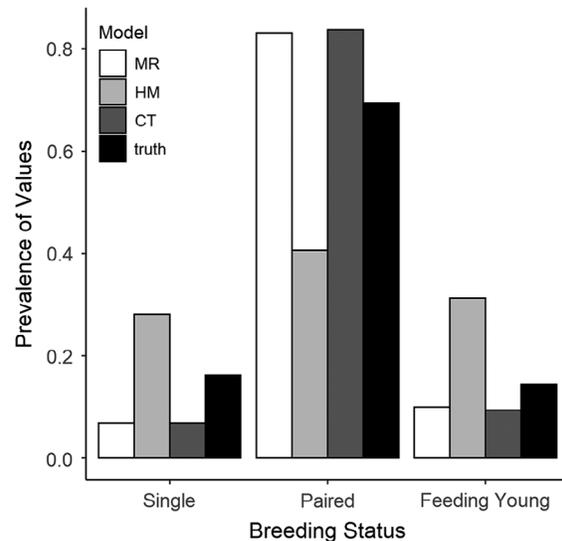

Fig. 3. Prevalence of predicted breeding statuses for the three model types (MR, multinomial logistic regression; HM, hierarchical model; and CT, classification tree) and the true breeding statuses.

daily variation in singing rate through the hierarchical series of conditional probability statements. This study provides an example of how to create a predictive model through statistical calibration for an indirect measurement of a biological state. Depending on the nature and scale of the research question or monitoring program, users could adapt the model to improve its predictive ability (i.e., sensitivity). Our work answers the call for more fundamental studies to better understand and represent the underlying mechanisms in indirect measurements in ecology (Lindenmayer and Likens 2011, Stephens et al. 2015).

The pattern we observed in how singing rate changes with breeding status in the OSFL is similar to patterns observed by Wright (1997) in an

OSFL population in Alaska. In both study areas, in different years, unpaired males sang at the highest rates; males who had paired and were engaged in initial breeding activities (i.e., nest building and incubating) sang less; and males feeding young rarely sang, and when they did, they sang few songs. Olive-sided Flycatchers in both studies also sang most around sunrise and much less as time since sunrise increased. This suggests that the song rate component of the hierarchical BSSR model can be used in different study areas for OSFL research and that time of day is an important song rate predictor. The other component of the hierarchical BSSR model, however, models the probability of each breeding status given the date or general species breeding timing (i.e., phenology) for that latitude and year. Although latitude was tested as a predictor in the breeding status component model, it was not significant in model selection. This result suggests limited regional variation in breeding phenology, contrary to our expectation that phenology would shift in our more northern study area. Environmental conditions during migration and at the breeding grounds can change breeding timing for a species among years, especially with the warming effects of climate change (Visser et al. 2004). Mean dates of OSFL pairing and of feeding young in 2016 in northern Alberta and the Northwest Territories were comparable to those reported from other OSFL populations, and phenology tended to not vary beyond a week between 1995 and 1996 from one study (Wright 1997). Although there may not be extreme variation in phenology between breeding regions or among years in OSFL, it may be important to verify breeding phenology for the region and year of interest for future applications of the model.

The BSSR model we produced is a simple version that can be used as a baseline on which to add parameters to improve predictive ability and/or apply to other vocal species. The singing behavior of many songbird species differs significantly between individuals with different breeding statuses (Radesäter et al. 1987, Gibbs and Wenny 1993, Stacier et al. 1996, Dussourd and Ritchison 2003). The modeling framework used in this study can likely be applied to other songbird species, but species-specific model calibrations would be required to select the most appropriate song metric and covariates. For example, instead of using song rate, length of song bout, time of first song, within-day variation in singing rates, song count conditional on at least one song, or a combination of these or other song metrics might provide better predictive ability. The BSSR model was constrained to using song rates recorded between OSFL spring arrival and late July when breeding OSFLs are feeding young. However, if the research objective was to predict breeding status after pairing (i.e., incubating eggs, feeding young, and fledging), the use of call rates instead of song rates may be a more precise indicator of nest status because calls represent activity at the nest (J. Hagelin and J. Wright, *unpublished data*). A wide range of unmeasured variables known to affect song rate in passerines may account for the 30–50% of unexplained variation in the BSSR model, such as density of conspecifics (Lampe and Espmark 1987), temperature (Gottlander 1987), within-individual variation (Radesäter et al. 1987, Robbins et al. 2009), and extra-pair copulation (Hasselquist et al. 1996). However, these factors require further investigation to assess their relative impact on BSSR predictive model performance.

The sensitivity values (i.e., true-positive rates) for the BSSR model were 69%, 50%, and 87% to predict single, paired, and feeding young, respectively. To our knowledge, no other studies have used calibration methods to predict breeding status from song rate, so we are unable to compare predictive ability with that from other models. However, we can compare the BSSR model predictive ability to that of other breeding bird reproductive indices. Vickery et al. (1992) designed a method to measure reproductive success, representing five statuses ranging from unpaired to fledged young, based on breeding behaviors. This index provided a reasonable measure of reproductive success for grassland songbirds compared to more intensive nest monitoring at the same study area (27% predicted fledged vs. 42% truly fledged; Vickery et al. 1992). When adapted to integrate nest monitoring with breeding behaviors for three forest breeding birds, the index provided correct breeding status predictions for 61–79% of the visits (Christoferson and Morrison 2001). Our hierarchical model had a similar predictive success, without the need for extensive nest searching and

behavioral observations, although our model is constrained to predict three breeding classes.

Monitoring song rate over a larger portion of the breeding season would improve certainty in predictions for individual birds. However, collection of song rate data by human observers on a fine temporal scale (i.e., daily vs. once per week) would take a large amount of time and likely be infeasible. A promising alternative method for collecting a larger amount of song rate data is using autonomous recording units (ARUs), which are increasing in popularity for bird research (Pankratz et al. 2017, Shonfield and Bayne 2017). There are three important advantages of using this technology to predict breeding status from song rate: (1) Daily acoustic surveys of a target location can be conducted for the entire breeding season; (2) large quantities of acoustic data can be processed using automatic recognition software; and (3) acoustic data can be permanently stored and thus may be reanalyzed later. The advantage connected to future reanalysis reflects the fact that automatic recognition software, used to detect species of interest efficiently, is still improving and future processing may improve detection rates on recordings. This technology is being applied over large spatial extents, and acoustic data are becoming readily available for many regions. For example, the Alberta Biodiversity Monitoring Institute (www.abmi.ca) has been monitoring breeding birds in Alberta since 2003 and breeding season-long recordings are available from across the province (Alberta Biodiversity Monitoring Institute 2011). Thus, large-scale demographic analyses based on temporal variation in song rates could be conducted for our focal species if the hierarchical model can be adapted for ARU-based song rates. Autonomous recording unit data have some challenges, however, primarily associated with imperfect detection probabilities related to bird movement away from the detection limit of the ARU. This would have to be accounted for in the modeling approach. The hierarchical model provides the framework to include such uncertainty and is an area of active investigation (E. Upham-Mills, et al., *unpublished manuscript*).

This study was the first attempt to predict a male songbird's breeding status using his singing rate, and our results provide a new method to monitor breeding status in a migratory songbird. We highlighted the importance of considering the calibration problem in ecological prediction modeling and demonstrated the advantage of using hierarchical modeling over conventional predictive model types (i.e., multinomial logistic regression and classification tree) to improve the sensitivity of predicting target classes. Future studies should aim at testing a similar approach to predict breeding status from song rates for other songbird species. We demonstrated that monitoring birdsong to infer songbird breeding status shows promise and warrants further investigation, especially if the model can be further developed for application with noninvasive ARU monitoring.

Acknowledgments

Literature Cited

Alberta Biodiversity Monitoring Institute. 2011. Monitoring activity report: lower Athabasca planning region 2010 season. Biodiversity Monitoring Report. Alberta Biodiversity Monitoring Institute, Edmonton, Alberta, Canada.

Ali, A., S. M. Shamsuddin, and A. L. Ralescu. 2015. Classification with class imbalance problem: a

review. International Journal of Advances in Soft Computing and its Applications 7:176–204.

Altman, B., and R. Sallabanks. 2012. Olive-sided Flycatcher (*Contopus cooperi*), version 2.0. *In* P. G. Rodewald, editor. The Birds of North America. Cornell Lab of Ornithology, Ithaca, New York, USA.

Anders, A. D., and M. R. Marshall. 2005. Increasing the accuracy of productivity and survival estimates in assembling landbird population status. Conservation Biology 19:66–74.

Arlot, S., and A. Celisse. 2010. A survey of cross-validation procedures for model selection. Statistics Surveys 4:40–79.

Armstrong, E. A. 1973. A study of bird song. Second edition. Dover Publications, Dover, New York, USA.

Berliner, L. M. 1996. Hierarchical Bayesian time series models. Pages 15–22 *in* K. Hanson and R. Silver, editors. Maximum entropy and Bayesian methods. Klewer, Norwell, Massachusetts, USA.

Bhapkar, V. P. 1966. A note on the equivalence of two test criteria for hypotheses in categorical data. Journal of the American Statistical Association 61:228–235.

Biondi, F., and K. Waikul. 2004. DENDROCLIM2002: a C ++ program for statistical calibration of climate signals in tree-ring chronologies. Computers & Geosciences 30:303–311.

Brieman, L., J. Friedman, C. J. Stone, and R. A. Olshen. 1984. Classification and regression trees. CRC Press, Boca Raton, Florida, USA.

Brooks, S. P., and A. Gelman. 1998. General methods for monitoring convergence of iterative simulations general methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 7:434–455.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: A practical information-theoretic approach. Second edition. Springer-Verlag, New York, New York, USA.

Christoferson, L. L., and M. L. Morrison. 2001. Integrating methods to determine breeding and nesting status of 3 western songbirds. Wildlife Society Bulletin 29:688–696.

Cressie, N., C. A. Calder, J. S. Clark, J. M. Ver Hoef, and C. K. Wikle. 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecological Applications 19:553–570.

De'Ath, G., and K. E. Fabricius. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81:3178–3192.

Dussourd, N., and G. Ritchison. 2003. Singing behavior of male yellow-breasted chats: repertoires, rates, reproductive success, and a comparison with other wood-warblers. Wilson Bulletin 115:52–57.

Ecosystem Classification Group. 2007. Ecological regions of the Northwest Territories – Taiga Plains, rev. 2009. Department of Environment and Natural Resources, Government of the Northwest Territories, Yellowknife, Northwest Territories, Canada.

Environment and Climate Change Canada. 2017. North American Breeding Bird Survey: Canadian Trends Website, Data-version 2015. Environment and Climate Change Canada, Gatineau, Quebec, Canada. https://wildlife-species.canada.ca/breeding-bird-survey-results

Gibbs, J. P., and D. G. Wenny. 1993. Song output as a population estimator: effect of male pairing status. Journal of Field Ornithology 64:316–322.

Gottlander, K. 1987. Variation in the song rate of the male Pied Flycatcher *Ficedula hypoleuca*: causes and consequences. Animal Behaviour 35:1037–1043.

Haché, S., P. Solymos, E. M. Bayne, T. Fontaine, S. G. Cumming, F. Schmiegelow, and D. Stralberg. 2014. Analyses to support critical habitat identification for Canada Warbler, Olive-sided Flycatcher, and Common Nighthawk, Final Report 2. Boreal Avian Modelling Project, Edmonton, Alberta, Canada.

Haché, S., M. Villard, and E. M. Bayne. 2013. Experimental evidence for an ideal free distribution in a breeding population of a territorial songbird. Ecology 94:861–869.

Hall, R. I., and J. P. Smol. 1992. A weighted-averaging regression and calibration model for inferring total phosphorus concentration from diatoms in British Columbia (Canada) lakes. Freshwater Biology 27:417–434.

Hamilton, J. D. 1986. A standard error for the estimated state vector of a state-space model. Journal of Econometrics 33:387–397.

Harris, I. 1989. Predictive fit for natural exponential families. Biometrika 76:675–684.

Hasselquist, D., S. Bensch, and T. von Schantz. 1996. Correlation between male song repertoire, extra-pair paternity and offspring survival in the Great Reed Warbler. Nature 381:229–232.

Holmes, R. T., T. W. Sherry, P. P. Marra, and K. E. Petit. 1992. Multiple brooding and productivity of a Neotropical migrant, the Black-throated blue warbler (*Dendroica caerulescens*), in an unfragmented temperate forest. Auk 109:321–333.

Hooten, M. B., and N. T. Hobbs. 2015. A guide to Bayesian model selection for ecologists. Ecological Monographs 85:3–28.

Hunt, A. R., E. M. Bayne, and S. Haché. 2017. Forestry and conspecifics influence Canada Warbler

(*Cardellina canadensis*) habitat use and reproductive activity in boreal Alberta, Canada. Condor 119:832–847.

Kéry, M., and J. A. Royle. 2016. Distribution, abundance, and species richness in ecology. Pages 3–18 *in* Applied hierarchical modeling in ecology. Elsevier, London, UK.

Kessel, B., and D. D. Gibson. 1978. Status and distribution of Alaska birds. *In* R. J. Raitt and J. P. Thompson, editors. Studies in avian biology no. 1. Allen Press, Lawrence, Kansas, USA.

Kuhn, M., and K. Johnson. 2016. Applied predictive modeling. Fifth edition. Springer Nature, New York, New York, USA.

Lampe, H. M., and Y. O. Espmark. 1987. Singing activity and song pattern of the Redwing *Turdus iliacus* during the breeding season. Ornis Scandinavica 18:179–185.

Lindenmayer, D. B., and G. E. Likens. 2011. Direct measurement versus surrogate indicator species for evaluating environmental change and biodiversity loss. Ecosystems 14:47–59.

Liu, W., D. E. Kroodsma, and K. Yasukawa. 2007. Dawn and daytime singing behavior of Chipping Sparrows (*Spizella passerina*). Auk 124:44–52.

Martin, T. E., C. Paine, C. J. Conway, W. M. Hochachka, P. Allen, and W. Jenkins. 1997. BBIRD (Breeding Biology Research & Monitoring Database) field protocol. Pages 1–64. Montana Cooperative Wildlife Research Unit, University of Montana, Missoula, Montana, USA.

McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12:153–157.

Northrup, J. M., and B. D. Gerber. 2018. A comment on priors for Bayesian occupancy models. PLOS ONE 13:1–14.

Olivas, P. C., S. F. Oberbauer, D. B. Clark, D. A. Clark, M. G. Ryan, J. J. O'Brien, and H. Ordoñez. 2013. Comparison of direct and indirect methods for assessing leaf area index across a tropical rain forest landscape. Agricultural and Forest Meteorology 177:110–116.

Osborne, C. 1991. Statistical calibration: a review. International Statistical Review 59:309–336.

Pankratz, R. F., S. Hache, P. Sólymos, and E. M. Bayne. 2017. Potential benefits of augmenting road-based breeding bird surveys with autonomous recordings. Avian Conservation and Ecology 12:18.

Pardieck, K. L., D. J. J. Ziolkowski, M.-A. R. Hudson, and K. Campbell. 2016. North American Breeding Bird Survey Dataset 1966–2015, version 2015.0. www.pwrc.usgs.gov/BBS/RawData/

Patterson, T. A., L. Thomas, C. Wilcox, O. Ovaskainen, and J. Matthiopoulos. 2008. State – space models of individual animal movement. Trends in Ecology & Evolution 23:87–94.

Peterson, R. T. 1980. A field guide to the birds east of the Rockies. Fourth edition. Houghton Mifflin, Boston, Massachusetts, USA.

Plummer, M., A. Stukalov, and M. Denwood. 2016. Package rjags: Bayesian graphical models using MCMC. https://cran.r-project.org/web/packages/rjags/rjags.pdf

R Core Team. 2017. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Radesäter, T., S. Jakobsson, N. Andbjer, A. Bylin, and K. Nyström. 1987. Song rate and pair formation in the Willow Warbler, *Phylloscopus trochilus*. Animal Behaviour 35:1645–1651.

Ripley, B., and W. Venables. 2016. Package nnet: feed-forward neural networks and multinomial log-linear models. http://www.stats.ox.ac.uk/pub/MASS4/

Robbins, M. B., R. D. S. Ny Ri, M. Pape, B. W. Benz, M. Papes, and M. B. Robbins. 2009. Song rates, mating status, and territory size of Cerulean warblers in Missouri Ozark riparian forest. Wilson Journal of Ornithology 121:283–289.

Robertson, B. A., J. J. Fontaine, and E. Loomis. 2009. Seasonal patterns of song structure variation in a suboscine passerine. Wilson Journal of Ornithology 121:815–818.

Robinson, P. W., Y. Tremblay, D. E. Crocker, M. A. Kappes, C. E. Kuhn, S. A. Shaffer, S. E. Simmons, and D. P. Costa. 2007. A comparison of indirect measures of feeding behaviour based on ARGOS tracking data. Deep Sea Research Part II: Topical Studies in Oceanography 54:356–368.

Salt, W. R., and J. R. Salt. 1976. The birds of Alberta with their ranges in Saskatchewan and Manitoba. Hurtig, Edmonton, Alberta, Canada.

Shonfield, J., and E. M. Bayne. 2017. Autonomous recording units in avian ecological research: current use and future applications. Avian Conservation and Ecology 12:14.

Spiegelhalter, D. J., A. Thomas, N. G. Best, and W. R. Gilks. 1995. BUGS: Bayesian inference using Gibbs sampling. Version 0.5. MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK.

Stacier, C. A., D. A. Spector, and A. G. Horn. 1996. The dawn chorus and other diel patterns in acoustic signaling. Pages 426–453 *in* D. E. Kroodsma and E. H. Miller, editors. Ecology and evolution of acoustic communication in birds. Cornell University Press, Ithaca, New York, USA.

Stacier, C. A., V. Ingalls, and T. W. Sherry. 2006. Singing behavior varies with breeding status of

American Redstarts (*Setophaga ruticilla*). Wilson Journal of Ornithology 118:439–451.

Stephens, P. A., N. Pettorelli, J. Barlow, M. J. Whittingham, and M. W. Cadotte. 2015. Management by proxy? The use of indices in applied ecology. Journal of Applied Ecology 52:1–6.

Strong, W. L., and K. R. Leggat. 1992. Ecoregions of Alberta. Alberta Forestry, Lands and Wildlife, Edmonton, Alberta, Canada.

ter Braak, C. J. F. 1995. Non-linear methods for multivariate statistical calibration and their use in palaeoecology: a comparison of inverse (k-nearest neighbours, partial least squares and weighted averaging partial least squares) and classical approaches. Chemometrics and Intelligent Laboratory Systems 28:165–180.

ter Braak, C. J. F., and L. G. Barendregt. 1986. Weighted averaging of species indicator values: its efficiency in environmental calibration. Mathematical Biosciences 78:57–72.

Therneau, T., B. Atkinson, and B. Ripley. 2018. Package rpart: recursive partitioning and regression trees. https://cran.r-project.org/web/packages/rpart/rpart.pdf

Thorpe, W. H. 1961. Bird-song. Cambridge University Press, Cambridge, UK.

Venables, W. N., and B. D. Ripley. 2002. Poisson and multinomial models. *In* J. Chambers, W. Eddy, W. Hardle, S. Sheather, and L. Tierney, editors. Modern applied statistics with S. Fourth edition. Springer Science + Business Media, New York, New York, USA.

Vickery, P. D., M. L. Hunter Jr., and J. V. Wells. 1992. Use of a new reproductive index to evaluate relationship between habitat quality and breeding success. Auk 109:697–705.

Visser, M. E., C. Both, and M. M. Lambrechts. 2004. Global climate change leads to mistimed avian reproduction. Advances in Ecological Research 35:89–110.

Wikle, C. K. 2003. Hierarchical Bayesian models for predicting the spread of ecological processes. Ecology 84:1382–1394.

Wright, J. M. 1997. Preliminary study of Olive-sided Flycatchers, July 1994 – April 1997. Final Research Report. Endangered species conservation fund federal aid studies SE-3-3, 4 and 5:34. Alaska Department of Fish and Game, Juneau, Alaska, USA.

Wu, X., et al. 2008. Top 10 algorithms in data mining. Knowledge and Information Systems 14:1–37.

SUPPORTING INFORMATION

Additional Supporting Information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/ecs2.3005/full