Bayesian hierarchical GAM to model BBS data

RESEARCH ARTICLE

Improved status and trend estimates from the North American Breeding Bird Survey using a Bayesian hierarchical generalized additive model

Adam C. Smith[1*] and Brandon P.M. Edwards [2]

[1]Canadian Wildlife Service, Environment Climate Change Canada, National Wildlife Research Centre, Ottawa Canada K1A 0H3.

ORCID: 0000-0002-2829-4843

[2]Department of Mathematics & Statistics, University of Guelph, Guelph Canada.

ORCID: 0000-0003-0865-3076

* Corresponding author: adam.smith2@canada.ca

ACKNOWLEDGEMENTS

We sincerely thank the thousands of U.S. and Canadian participants who annually perform and coordinate the North American Breeding Bird Survey. We also wish to acknowledge Courtney Amundson for sharing some code on similar models and John Sauer and Bill Link for sharing code that helped with the cross-validations and for many spirited, collegial discussions that have informed this work. We also thank the many biologists within the Canadian Wildlife Service and other users of the BBS status and trend estimates whose insightful questions and suggestions motived much of this work, including Charles Francis, Marie-Anne Hudson, Veronica Aponte, Marcel Gahbauer, Pete Blancher, and Ken Rosenberg.

Data Depository: R scripts to download the BBS data and to perform the analyses in this paper and are archived at www.github.com/AdamCSmithCWS/GAM_Paper_Script

Funding Statement: This work was supported by operating funds from Environment and Climate Change Canada

Ethics Statement: This research was conducted in compliance with the Environment and Climate Change Canada Values and Ethics Code.

Author Contributions: ACS conceived the ideas and designed methodology; ACS and BPME analyzed the data; ACS led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ABSTRACT

The status and trend estimates derived from the North American Breeding Bird Survey (BBS), are critical sources of information for bird conservation. However, many of the varied uses of these estimates are poorly supported by the current standard model. For example, inferences about population recovery require modeling approaches that are more sensitive to changes in the rates of population change through time and population cycles. In addition, regional status assessments would benefit from models that allow for the sharing of information across the species' range. Here we describe Bayesian hierarchical generalized additive models (GAM) that fit these criteria, generating status and trend estimates optimized for many common uses related to conservation assessments. We demonstrate the models and their benefits using data for Barn Swallow (*Hirundo rustica*), Wood Thrush (*Hylocichla mustelina*), Carolina Wren (*Thryothorus ludovicianus*), and a selection of other species, and we run a full cross-validation of the GAMs against two other BBS models to compare predictive fit. The GAMs have better predictive fit than the standard model for all species studied here, and better or comparable predictive fit compared to an alternative first difference model. In addition, one version of the GAM described here (GAMYE) estimates a population trajectory that can be decomposed into a smooth component and the annual fluctuations around that smooth. This decomposition also allows trend estimates based only on the smooth component, which are more stable between years and are therefore more useful for trend-based status assessments, such as those by the IUCN. It also allows for the easy customization of a BBS model to incorporate covariates that influence the smooth component separately from those that influence annual fluctuations (e.g., climate cycles vs annual precipitation). This GAMYE model is a broadly useful model for the BBS and other

long-term surveys, because of its flexibility, its decomposition, and the hierarchical structure that shares information among regions.

Keywords: Bayesian, Breeding bird survey, Cross validation, Generalized additive model, Population change, Status and trend estimates

LAY SUMMARY

- We describe an improved way to estimate population status and trends from the North American Breeding Bird Survey data, using a Bayesian hierarchical generalized additive mixed-model.
- The model generates estimates with better predictive accuracy than the previous model.
- Status and trend estimates from the model are more broadly useful for a wide range of common conservation applications

INTRODUCTION

Estimates of population change derived from the North American Breeding Bird Survey are a keystone of avian conservation in North America. Using these data, the Canadian Wildlife Service (CWS, a division of Environment and Climate Change Canada) and the United States Geological Survey (USGS) produce national and regional status and trend estimates (estimates of annual relative abundance and rates of change in abundance, respectively) for 300-400 species of birds (Smith et al. 2019, Sauer et al. 2014). These estimates are derived from models designed to account for some of the sampling imperfections inherent to an international, long-term field survey, such as which sites or routes are surveyed in a given year and variability among observers (Sauer and Link 2011, Smith et al. 2014). Producing these estimates requires significant analytical expertise, time, and computing resources, but they are used by many conservation organizations and researchers to visualize, analyze, and assess the population status

of over 400 species of birds (e.g., Rosenberg et al. 2017, NABCI Canada 2019, Rosenberg et al. 2019).

While the estimates of status and trend from the BBS serve many different purposes, not all uses are well supported by the standard models, and so there is a need for alternative models and for a continual evolution of the modeling. Different conservation-based uses of the BBS status and trend estimates relate to different aspects of population change, including long-term trends for overall status (Partners in Flight, 2019), short-term trends to assess extinction-risk (IUCN 2019), changes in population trends to assess species recovery (Environment Climate Change Canada, 2016), or annual fluctuations (Wilson et al., 2018). Each one of these uses relies on different parameters and or spatial and temporal variations in those parameters, and no single model can estimate all parameters equally well. This is not a criticism; it is true of any single model. For example, the standard model used since approximately 2011, estimates population change using slope-parameters and random year-effects in a Bayesian hierarchical framework (Sauer and Link 2011, Smith et al. 2014). These slope and year-effects are well suited to estimating annual fluctuations around a continuous long-term change, but the model tends to be conservative when it comes to estimating changes in a species' population trend, or population cycles (Fewster et al. 2000, Smith et al. 2015). Similarly, short-term trends (e.g., the last 10-years of the time-series) derived from the standard models incorporate information from the entire time-series (i.e., the slope component of the model). For many purposes, this is a reasonable and useful assumption, which guards against extreme and imprecise fluctuations in short-term trends. However, for assessing changes in trends of a once-declining species, such as the recovery of a species at risk (Environment Climate Change Canada, 2016), this feature of the model is problematic.

Generalized Additive Models (GAM, Wood 2017) provide a flexible framework for tracking changes in populations over time, without any assumptions about a particular pattern in population change (Fewster et al., 2000, Knape 2016). The semi-parametric smooths can fit almost any shape of population trajectory, including stable populations, constant rates of increase or decrease, cycles of varying frequency and amplitude, or abrupt change points in population trends (Wood 2017). Furthermore, the addition of new data in subsequent years has little or no influence on estimates of population change in the earlier portions of the time-series. By contrast, the slope parameter in the standard models estimates a constant rate of population change across the entire time-series, effectively assuming that there is some consistent rate of change. As a result, estimates of the rate of a species population change in the early portion of the time series (e.g., during the 1970s or 80s) will change in response to the addition of contemporary data and recent rates of population change. The random walk structure of a first-difference model (Link et al. 2017) assumes that the population in each year is similar to the population in the previous year—a biologically reasonable assumption. However, if data are sparse in a given region and year, the population trends are shrunk towards zero, effectively assuming a stable population. This stability-prior is usually overwhelmed by the data, but the user of published trend estimates has no clear way to discern its influence.

GAMs also provide a useful framework for sharing information on the shape and rate of population change across a species range. The GAM smoothing parameters can be estimated as random effects within geographic strata, thus allowing the model to share information on the shape of a species population trajectory across a species range. In the terminology of Pedersen et al. 2019, this hierarchical structure on the GAM parameters would make our model a “HGAM” (Hierarchical Generalized Additive Model). However, it also includes random effects for

parameters not included in the smooth and could therefore be referred to as a GAMM (Generalized Additive Mixed Model), in the terminology of Wood 2017. In the standard model, the slope parameters can be estimated as random effects and share information among strata, which improves estimates of trend for relatively data-sparse regions (Link et al. 2017, Smith et al. 2019). Although recent work has shown that the standard model is, for many species, out-performed by a first-difference model (Link et al. 2020), the population change components of the first-difference model (Link et al. 2017), include no way to share information on population change in space and so population trajectories are estimated independently among strata.

The inherently smooth temporal patterns generated by GAMs are well suited to particularly common conservation uses, such as assessments of trends in populations from any portion of a time-series, as well as assessments of the rate of change in the trends over time. For example, the population trend criteria of the IUCN (IUCN 2019) or Canada's national assessments by the Committee on the Status of Endangered Wildlife in Canada (COSEWIC) are based on rates of change over 3 generations. For most bird species monitored by the BBS, this 3-generation time is approximately the same as the 10-year, short-term trends produced by the CWS and USGS analyses. Because of the annual fluctuations estimated by the standard model, these short-term trends can fluctuate from year to year, complicating the quantitative assessment of a species trend in comparison to the thresholds. Species trends may surpass the threshold in one year, but not in the next. The same end-point comparisons on estimates from a GAM will change much more gradually over time, and be much less dependent on the particular year in which a species is assessed.

In this paper, we describe a status and trend model that uses a hierarchical GAM to estimate the relative abundance trajectory of bird populations, using data from the BBS. This model allows

for the sharing of information about a species' population trajectory among geographic strata and for the decomposition of long- and medium-term population changes from annual fluctuations. We also compare the fit of the GAM, and a GAM-version that includes random year-effects (Knape et al. 2016), to the fit of two alternative models for the BBS (Sauer and Link 2011, Smith et al. 2015, Link and Sauer 2019).

METHODS

Overview

We designed a Bayesian hierarchical model for estimating status and trends from the North American Breeding Bird Survey (BBS) that uses a Generalized Additive Model (GAM) smooth to estimate the medium- and long-term temporal components of a species population trajectory. In the model, the GAM smooths are treated as random-effects within the geographic strata (the spatial units of the predictions, intersections of Bird Conservation Regions and province/state/territory boundaries), so that information is shared on the shape of a species' population trajectory among sub-regions of the species' range. In comparison to the non-Bayesian hierarchical GAMs (HGAM) in Pedersen et al. 2019, our model is most similar to the "GS" model that has a global smoother and group-level smoothers with a similar degree of wiggliness. We applied two versions of the GAM: one in which the GAM smooth was the only component modeling changes in abundance over time (GAM), and another in which random year effects were also estimated to allow for single-year departures from the GAM smooth (GAMYE, which is conceptually similar to the model described in Knape 2016).

For a selection of species, we compared estimates and predictive accuracy of our two models using the GAM smooth, against two alternative models that have been used to analyze the BBS.

Our models use the counts of individual birds observed on BBS routes in a given year by a particular observer. The BBS routes are roadside survey-routes that include 50 stops at which a 3-minute point count is conducted, once annually, during the peak of the breeding season (Robbins et al. 1986). The four statistical models differed only in the parameters used to model changes in species relative abundance over time. We used 15-fold cross validation (Burman 1983) to estimate the observation-level, out-of-sample predictive accuracy of all four models (Link et al. 2019, Vehtari et al. 2017). We compared the overall predictive accuracy among the models, and we explored the spatial and temporal variation in predictive accuracy in depth.

In addition to the GAMs, we compared four alternative BBS models, all of which have the same basic structure:

$$\log\left(\lambda_{s,j,t}\right) = \theta_s + \Delta_s(t) + \eta I[j,t] + \omega_j + \varepsilon_{s,j,t}$$

The models treat the observed BBS counts as overdispersed Poisson random variables, with mean $\lambda_{s,j,t}$ (i.e., geographic stratum s, observer and route combination j, and year t). The means are log-linear functions of stratum-specific intercepts (θ_s), observer-route effects (ω_j), first-year startup effects for a given observer ($\eta I[j,t]$) a count-level random effect to model overdispersion ($\varepsilon_{s,j,t}$), and a temporal component estimated using a function of year, which varies across the four models ($\Delta_s(t)$). The models here only varied in their temporal components ($\Delta_s(t)$). We set the priors following Link et al. (2017) and Smith et al. (2019), and we used the heavy-tailed, t-distribution to model the count-level extra-Poisson dispersion (Link et al. 2019).

Bayesian hierarchical GAMs

GAM. The main temporal component $\Delta_s(t)$ in the GAM was modeled with a semi-parametric smooth, estimated following Crainiceanu et al (2005) as

$$\Delta_{\mathrm{s}}^{\mathrm{GAM}}(t) = \sum_{k=1}^{K} \beta_{s,k} \chi_{t,k}$$

where K is the number of knots, $\chi_{t,k}$ is the year t and kth entry in the design matrix X (defined below), and $\beta_{i,.}$ is the K-length vector of parameters that control the shape of the trajectory in stratum s. Each

$\beta_{s,k}$ is estimated as a random effect, centered on a mean across all strata (a hyperparameter B_k)

$$\beta_{s,k} \sim Normal\left(\mathrm{B}_k, \sigma_\beta^2\right)$$

and

$$\mathrm{B}_K \sim Normal(\mathbf{0}, \sigma_{\mathrm{B}}^2)$$

where the variance σ_{B}^2 controls the shrinkage towards a first-degree polynomial. That is, this random effect variance term σ_{B}^2 acts as the complexity penalty, shrinking the wiggliness of the mean trajectory towards a straight and flat line). In combination, these variance parameters ($\sigma_\beta^2, \sigma_{\mathrm{B}}^2$) control the complexity penalty of the species trajectories and the variation in pattern and complexity among strata and were given the following priors, following advice in Crainiceanu et al (2005):

$$\sigma_\beta^2 \sim \frac{1}{gamma(2, 0.2)}$$

$$\sigma_{\mathrm{B}}^2 \sim \frac{1}{gamma(10^{-2}, 10^{-4})}$$

These prior parameters were chosen as a balance between ensuring that the priors are sufficiently vague and yet constrained to realistic values given the scale of the basis function. We have so far had good results across a wide range of species using these priors, but they are an area of ongoing research and users of the package *bbsBayes* can access the JAGS model descriptions to explore alternative priors.

The design matrix for the smoothing function (X) has a row for each year, and a column for each of K knots. The GAM smooth represented a 3rd-degree polynomial spline: $\chi_{t,k} = |t' - t'_k|^3$ and was calculated in R, following Crainiceanu et al (2005). Here, we have used 13 knots (K = 13), across the 53 year time-series of the BBS (1966-2018), following the default setting in the R-package *bbsBayes* (Edwards and Smith 2020), which is to add one knot for every 4 years in the time-series. With this number of knots, we have found that the 53-year trajectories are sufficiently flexible to capture all but the shortest-term variation (e.g., variation on the scale of 3-53 years, but not annual fluctuations).

GAMYE. The GAMYE was identical to the GAM, with the addition of random year effects ($\gamma_{t,s}$) estimated following Sauer and Link (2011) and Smith et al. (2015), as

$$\gamma_{t,s} \sim Normal\left(\mathbf{0}, \sigma_{\gamma,s}^2\right)$$

where $\sigma_{\gamma,s}^2$ are stratum-specific variances. Thus, the temporal component for the GAMYE is given by

$$\Delta_s^{GAMYE}(t) = \sum_{k=1}^{K} \beta_{s,k}\chi_{t,k} + \gamma_{t,s}$$

Alternative models

For a selection of species, we compared the predictions and predictive accuracy of the two GAMs against two alternative models previously used for the BBS.

SLOPE. The SLOPE model includes a log-linear slope parameter and random year-effects to model species trajectories. It is a linear-trend model currently used by both the CWS (Smith et al. 2014) and the USGS (Sauer et al. 2017) as an omnibus model to supply status and trend estimates from the BBS (essentially the same as model SH, the Slope model with Heavy-tailed error in Link et al 2017). The main temporal component in the SLOPE model is

$$\Delta_s^{SLOPE}(t) = \beta_s * (t - t_{mid}) + \gamma_{t,s}$$

DIFFERENCE. The first-difference model (DIFFERENCE) is based on a model described in Link and Sauer (2015) and models the main temporal component as

$$\Delta_s^{DIFFERENCE}(t) = \gamma_{t,s} = N\left(\gamma_{t-1,s}, \sigma_{\gamma_s}^2\right)$$

The DIFFERENCE model includes year-effects that follow a random walk from the first year of the time-series, by modeling the first-order differences between years as random effects with mean zero and an estimated variance.

All analyses in this paper were conducted in R (R Core Team, 2019), using JAGS to implement the Bayesian analyses (Plummer 2003), and an R-package *bbsBayes* (Edwards and Smith 2020) to access the BBS data and run all of the models used here. The graphs relied heavily on the package *ggplot2* (Wickham 2016). BUGS-language descriptions of the GAM and GAMYE, as

well as all the code and data used to produce the analyses in this study, are archived online (see Data Depository in Acknowledgements). In addition, all of the models used here can be applied to the BBS data using the R-package "*bbsBayes*" (Edwards and Smith 2020).

Cross-validation

We used a temporally and spatially stratified v-fold cross-validation (Burman 1983, often termed "k-fold", but here we use Berman's original "v-fold" to distinguish it from "k" which is often used to describe the number of knots in a GAM) with v = 15, where we held-out random sets of counts, stratified across all years and strata so that each of the v-folds included some observations from every combination of strata and years. We chose this approach over a leave-one-out cross-validation (loocv) approach using a random subset of counts (e.g., Link et al. 2019) because we wanted to assess the predictive success across all counts in the dataset, explore the temporal and spatial patterns in predictive success, and a full loocv is not practical for computational reasons (see Link et al. 2017). We followed a similar procedure to that outlined in Link et al. (2017) to implement the cross-validation in a parallel computing environment, using the R-package foreach (Microsoft and Weston 2019). We did not calculate WAIC because previous work has shown that WAIC does not approximate loocv well for the BBS data (Link et al. 2017, Link et al. 2019).

We used the estimated log predictive density ($\mathrm{elpd}_{i,M}$) to compare the observation-level, out-of-sample predictive success of all four models (Link et al. 2019, Vehtari et al. 2017). For a given model M, elpd is the estimated log posterior probability for each observation i, for the model M fit to all data except those in the set v that includes i ($Y_{-k,i\in k}$). That is,

$$\mathrm{elpd}_{i,M} = \log\left(f_M\left(Y_i | Y_{-k,i\in k}, X_i\right)\right)$$

Larger values of elpd indicate better predictive success, that is a higher probability of the observed data given the model M, the covariate vector (X_i), and all of the data used to fit the model ($Y_{-K,i\in K}$).

We have not summed elpd values to generate BPIC values (Link et al. 2019); rather, we have compared model-based estimates of mean difference in elpd between pairs of models. To compare the prediction error between pairs of models, we calculated the difference in the elpd of each observed count (Y_i) under models 1 and 2, as $\delta_{i,M1-M2}^{elpd} = \log\left(f_1(Y_i|Y_{-i}, X_i)\right) - \log\left(f_2(Y_i|Y_{-i}, X_i)\right)$, so that positive values of $\delta_{i,M1-M2}^{elpd}$ indicate more support for model 1. We analysed these δ_i^{elpd} values using an additional Bayesian hierarchical model to account for the imbalances in the BBS-data among years and regions, and the inherent uncertainty associated with any cross-validation statistic (Vehtari et al. 2017, and Link et al. 2017). This model treated the elpd differences for a count from a given year t and stratum s ($\delta_{i,s,t}^{elpd}$) as having a t-distribution with an estimated variance (σ_δ^2) and degrees of freedom (ν). That is,

$$\delta_{i,s,t}^{elpd} = t(\mu_i, \sigma_\Delta^2, \nu)$$

$$\mu_i = \phi + \psi_s + \psi_t$$

From the model, ϕ was our estimate of the overall comparison of the mean difference in predictive fit for Model 1 – Model 2 ($\delta_{M1-M2}^{elpd} = \phi$), $\phi + \psi_s$ was the estimate of the mean difference in stratum s, and $\phi + \psi_t$ was the estimated difference in year t. We used this robust estimation approach, instead of the z-score approach used by Link and Sauer (2020) because of the extremely heavy tails in the distribution of the δ_i^{elpd} values. Given these heavy tails, a statistical analysis assuming a normal distribution would give an inappropriately large weight to

a few extremely poorly predicted counts (Gelman et al. 2014). In essence, our model is simply a "robust" version of the z-score approach (Lange et al. 1989) with the added hierarchical parameters to account for the spatial and temporal imbalance in the BBS data.

Trends and population trajectories

For all models, we used the same definition of trend following Sauer and Link (2011) and Smith et al. (2015); that is, an interval-specific geometric mean of proportional changes in population size, expressed as a percentage. Thus, the trend estimate for the interval from year a (t_a) through year b (t_b) is given by

$$\mathrm{R}_{\mathrm{a:b}} = 100 * \left(\left(\frac{N_{t_a}}{N_{t_b}} \right)^{\frac{1}{t_a - t_b}} - 1 \right)$$

where $N_{\cdot}$ represents the annual index of abundance in a given year (see below). Because this estimate of trend only considers the annual abundance estimates in the years at either end of the trend period, we refer to this estimate as an end-point trend. For the GAMYE model, we decomposed the trajectory (i.e., the series of annual indices of abundance) into long- and medium-term components represented by the GAM smooth and annual fluctuations represented by the random year-effects. This decomposition allowed us to estimate two kinds of trend estimates: $\mathrm{R}_{\mathrm{a:b}}$ that include all aspects of the trajectory, and $\mathrm{R}'_{\mathrm{a:b}}$ that removes the annual fluctuations, including only the GAM smooth components.

Population trajectories are the collection of annual indices of relative abundance across the time series. For all the models here, we calculated these annual indices for each year t and stratum s following Smith et al. (2019) as

$$N_{s,t} = z_s * \frac{\sum_{j \in J_s} e^{\mathrm{A}_{s,t} + \omega_j + 0.5 * \sigma_\varepsilon^2}}{n_{Js}}$$

where each $N_{s,t}$ are exponentiated sums of the relevant components of the model ($\mathrm{A}_{s,t}$), observer-route effects (ω_j), and count-level extra-Poisson variance ($0.5 * \sigma_\varepsilon^2$), averaged over count-scale predictions across all of the n_{Js} observer-routes j in the set of observer-route combinations in stratum s (J_s), and then multiplied by a correction factor for the proportion of routes in the stratum on which the species has been observed (z_s). This is conceptually similar to the approach described in Sauer and Link (2011) and Smith et al. (2015); for a discussion on the differences, refer to the Appendix.

For the GAMYE model, we calculated two versions of the species trajectory (N_s): one that included the annual variation in the trajectory,

$$N_{s,t} = z_s * \frac{\sum_{j \in J_s} e^{\mathrm{A}_{s,t} + \omega_j + 0.5 * \sigma_\varepsilon^2}}{n_{Js}}$$

$$\mathrm{A}_{s,t} = \alpha_s + f_s(t) + \gamma_{s,t}$$

and a second that excluded the annual variations, including only the smoothing components of the GAM to estimate the time-series,

$$Ng_{s,t} = z_s * \frac{\sum_{j \in J_s} e^{\mathrm{Ag}_{s,t} + \omega_j + 0.5 * \sigma_\varepsilon^2}}{n_{Js}}$$

$$\mathrm{Ag}_{s,t} = \alpha_s + f_s(t)$$

We calculated population trajectories and trends from the GAMYE model using both sets of annual indices ($N_{s,t}$ and $Ng_{s,t}$). When comparing predictions against the other models, we use the $\mathrm{N}_{s,t}$ values to plot and compare the population trajectories (i.e., including the year-effects), and the $\mathrm{Ng}_{s,t}$ values to calculate the trends (i.e., removing the year-effect fluctuations).

RESULTS

Model predictions:

Population trajectories from the GAM and GAMYE are very similar. Both models suggest that Barn Swallow populations increased from the start of the survey to approximately the early 1980s, compared to the standard model predictions that show a relatively steady rate of decline (Figure 1). The trajectories for all species from the GAMs and less linear overall than the SLOPE model trajectories and tend to better track the nonlinear patterns in the raw data (Supplemental Materials Figure S1). GAM and GAMYE trajectories vary a great deal among the strata, particularly in the magnitude and direction of the long-term change (Figure 2). However, there are also many similarities among the strata, in the non-linear patterns (e.g., change points) that are evident in the continental mean trajectory (Figure 2 and Supplemental Materials Figure S2). Figure 3 shows the estimate trajectories for Barn Swallow in the 6 strata that make up BCR 23 from the GAMYE and SLOPE model. The GAMYE estimates suggest that the species' populations increased in the early portion of the time series in all of the strata, but the estimates from the SLOPE model only show that increase in the stratum with the most data (i.e., the most stacked grey dots along the x-axis indicating the number of BBS routes contributing data in each year, US-WI-23). In the other strata with fewer data the trajectories are strongly linear. The

cross-validation results suggest that the GAMYE is preferred over the SLOPE model for Barn Swallow and for the other species considered here (Figure 4)

The decomposed trajectories from the GAMYE allow us to calculate trends from the smooth and plot trajectories that show the annual fluctuations. For example, the smooth trajectory for the Carolina Wren captures the general patterns of increases and decreases well, while the full trajectory also shows the sharp population crash associated with the extreme winter in 1976 (Figure 5). Calculating trends from the smooth component generates short-term estimates that vary less from year to year for species with relatively strong annual fluctuations (Figure 6). For example, Figure 7 shows the series of short-term (10-year) trend estimates for Wood Thrush in Canada, from the GAMYE including the year-effects, the GAMYE from just the smooth component, and the SLOPE model used since 2011. In this example, the 10-year trend estimate from the GAMYE with the year-effects and the SLOPE model both cross the IUCN trend threshold criterion for Threatened (IUCN 2019) at least once in the last 12 years. In fact, the SLOPE model estimates flip from one side of the threshold to the other three times, and were below the threshold in 2011, when the species' status was assessed in Canada (COSEWIC 2012). By contrast, a trend calculated from the decomposed GAMYE model using only the smooth component (GAMYE – Smooth Only in Figure 7) suggests that the species is decreasing relatively steadily and gives a consistent signal about the rate of decline in comparison to the IUCN criterion.

Comparison to other models - Cross-validation

For the Barn Swallow, the SLOPE model trajectories are noticeably different from the other three models (Figure 8). The predictive fit comparisons suggest that the estimates from the three models that show an initial increase in Barn Swallow populations up to about the early 1980s are

preferable to the SLOPE predictions that show a more constant rate of decline (Figures 4 and 9). In contrast, the GAMYE and the DIFFERENCE model had very similar estimates (Figure 8) and relatively similar predictive fit values for most species (Figures 4 and 9, and Supplemental Material Figure S3).

The preferred model from the pairwise predictive fit comparisons depended on the species and varied in time and space (Figures 4, 9, 10 and Supplemental Material Figure S3 and S5). The contrast between GAMYE and DIFFERENCE for Barn Swallow provide a useful example: Depending on the year or the region of the continent, either the GAMYE or the DIFFERENCE model was the preferred model, but overall, and in almost all regions and years, the 95% CI of the mean difference in fit between GAMYE and DIFFERENCE overlapped 0 (Figures 4, 9 and 10). Barn Swallow provide an example of this complexity, the GAMYE model has generally higher predictive fit during the first 5 years of the time-series, but then the DIFFERENCE model is preferred between approximately 1975 and 1983. The geographic variation in predictive fit is similarly complex. In the Eastern parts of the Barn Swallow's range, the GAMYE model generally outperforms the DIFFERENCE model, whereas the reverse is generally true in the remainder of the species' range (Figure 10). Although the mapped colours only represent the point-estimates, they suggest an interesting spatial pattern in the predictive fit of these two models for this species. Many of species considered here show similarly complex temporal patterns of predictive fit (Supplemental Material Figures S5 and S6).

DISCUSSION

Using Bayesian hierarchical semi-parametric GAM smooths to model time series of population abundance with the North American Breeding Bird Survey generates more useful estimates of population trajectories and trends than other models and has comparable or better out of sample

predictive accuracy. The flexibility of the GAM smoothing structure to model long- and medium-term temporal patterns, and the optional addition of random year-effects to model annual fluctuations, allow it to model a wide range of temporal patterns within a single base-model (Fewster et al. 2000, Wood 2017). We fit the smooth components as random effects, to share information across geographic strata within a species' range, and to improve the estimates of population trajectories for data-sparse regions (Pedersen et al. 2018). For almost all species included here, the two GAM-based models clearly out-performed the standard model (SLOPE) used for the CWS and USGS analyses since 2011 (Sauer and Link 2011, Smith et al. 2014), and showed similar out of sample predictive accuracy as a first-difference, random-walk trajectory model (Link et al. 2020). On a practical note, the GAM-based models required approximately 30% more time than the SLOPE model to generate a similar number of posterior samples, but given the 53 years of effort to collect the data, we suggest this is a small price to pay for more useful status and trend estimates.

The decomposition of the estimated population trajectory into the smooth and year-effect components is a unique feature of the GAMYE and particularly useful for conservation applications. It allows the user to estimate and visualize separate trends and trajectories that include or exclude the annual fluctuations (Knape 2016). This allows the estimates to suit a range of conservation and management applications that rely on visualizing and estimating multiple aspects of population change. For example, the smoothed population trajectories capture the medium- and long-term changes in populations that are most relevant to broad-scale, multi-species assessments like the "State of the Birds" reports (NABCI-Canada 2019) where the annual fluctuations of a given species are effectively noise against the signal of community level change over the past 50 years (e.g., Rosenberg et al. 2019). Similarly, estimates of population

trends (interval-specific, rates of annual change) derived from the smooth component are responsive to medium-term changes and so can be used to identify change points in trends such as the recovery of Species at Risk (Environment Climate Change Canada 2016).

Trends derived from the smooth component of the GAMYE are also much less likely to fluctuate from year to year and therefore more reliable for use in species at risk status assessments. In many status assessments, such as those by IUCN and COSEWIC, population declines beyond a particular threshold rate (e.g.,) can trigger large investments of resources related to policy and conservation actions. For example, in both the IUCN red-listing and Canada's COSEWIC species at risk assessments (IUCN 2019) estimated declines greater than 30% over three generations is one criteria that results in a "Threatened" designation. If the estimated rate of population decline is strongly dependent on the particular year in which a species is assessed, there is an increased risk of inaccurate assessments, leading to failures to protect species, or inefficient investments of conservation resources. Of course, the full assessments of species' status are sophisticated processes that consider far more than just a single trend estimate. However, the example in Figure 6 raises the question of whether Wood Thrush would have been assessed as Threatened in Canada if the relevant trend had not been estimated in 2011 (COSEWIC 2012).

In some conservation or scientific uses of the BBS estimates, the annual fluctuations may be important components of the trajectory (e.g., winter-related mortality of Carolina Wrens), and so the full trajectory that includes both components from the GAMYE is most useful. This comprehensive estimate of a species' population trajectory is likely the best approach for the official presentation of a time series. At a glance, managers, conservation professionals, and researchers can glean information about fluctuations that might relate to annual covariates such

as precipitation, wintering ground conditions, or cone-crop cycles. The GAMYE structure allows an agency like the CWS to provide estimates in multiple versions (e.g., full trajectories, smoothed trajectories, trends), drawn from a coherent model, to suit a wide range of conservation applications, and to produce them in an efficient way. For example, there are situations where the ability for a user to access a ready-made separation of the yearly fluctuations from the underlying smooth could be helpful in the initial formulation of an ecological hypothesis. In addition, for custom analyses (Edwards and Smith 2020) a researcher can modify the basic GAMYE model to include annual covariates on the yearly fluctuations (e.g., spruce cone cycles) and other covariates on the smooth component (e.g., climate cycles).

Predictive accuracy

Overall, the cross-validation comparisons generally support the GAMYE, GAM, or DIFFERENCE model over the SLOPE model for the species considered here. For Barn Swallow, the overall difference in predictive fit, and particularly the increasing predictive error of the SLOPE model in the earliest years, strongly suggests that in the period between the start of the BBS (1966) and approximately 1983 (Smith et al. 2015), Barn Swallow populations increased. All models agree, however, that since the mid-1980's populations have decreased.

Using all data in our cross-validations allowed us to explore the spatial and temporal variation in fit, and to compare the fit across all data used in the model. Estimates of predictive fit from a random selection of BBS counts are biased by the strong spatial and temporal dependencies in the BBS data (Roberts et al. 2017). However, because our folds were identical across models, and we modeled the differences in fit with an additional hierarchical model that accounted for repeated measures among strata and years, we are reasonably confident that relative-fit assessments are unbiased.

The overall predictive fit assessments provided some guidance on model selection for the species here, but not in all cases. The SLOPE model compared poorly against most other models in the overall assessment, similar to Link et al. 2020. However, among the other three models, many of the overall comparisons failed to clearly support one model, even in cases where the predicted population trajectories suggested very different patterns of population change (e.g,. Cooper's Hawk in Supplemental Material Figure S5). For a given species the best model varied among years and strata. These temporal and spatial patterns in predictive fit complicate the selection among models, given the varied uses of the BBS status and trend estimates (Rosenberg et al. 2017).

In general, estimates of predictive accuracy are one aspect of a thoughtful model building and assessment process, but are insufficient on their own (Gelman et al. 2013 pg 180, Burnham and Anderson 2002 pg 16). This is particularly true if there is little or no clear difference in overall predictive accuracy, but important differences in model predictions. For example, the overall cross validation results do not distinguish between the SLOPE and GAMYE for Cooper's Hawk, and yet predictions are very different in the first and last few years of the time series (Figure 4). In contrast, the overall cross validation results for Carolina Wren suggest the DIFFERENCE model is preferred over the GAMYE, and yet the trajectories are almost identical (Figure 4 and Supplemental Material Figure S5). Predictive accuracy is also complicated when robust predictions are required for years or regions with relatively few data against which predictions can be assessed (e.g., the earlier years of the BBS). We agree with Link et al. (2020) that we should not select models based on a particular pattern in the results. In fact, the necessary subjective process occurs before any quantitative analyses (Burnham and Anderson 2002), and relies on "careful thinking" to balance the objectives; the model; and the data (Chatfield 1995).

The careful thinking required to select a BBS model or to interpret the BBS status and trend estimates, is to consider the consequences of the potential conflicts between the model structures ("constraints on the model parameters" sensu Chatfield 1995) and the objectives of the use of the modeled estimates. We suggest that the GAMYE's strong cross-validation performance and its flexible assumptions that match the most common uses of the BBS status and trends estimates, make it a particularly useful model for the sort of omnibus analysis conducted by the CWS and other agencies.

REFERENCES CITED

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.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference (with discussion). Journal of the Royal Statistical Society (London), Series A 158:419–466. doi:10.2307/2983440

COSEWIC. (2012). COSEWIC assessment and status report on the Wood Thrush *Hylocichla mustelina* in Canada. Committee on the Status of Endangered Wildlife in Canada. Ottawa.

Crainiceanu CM, Ruppert D, Wand MP (2005). Bayesian Analysis for Penalized Spline Regression Using WinBUGS. Journal of Statistical Software, 14 (14). doi:10.18637/jss.v014.i14

Duan, N. (1983). Smearing Estimate: A Nonparametric Retransformation Method. Journal of the American Statistical Association, Vol. 78, No. 383, pp. 605-610. doi:10.2307/2288126

Edwards, B.P.M. and A.C. Smith (2020). bbsBayes v2.1.0 (Version 2.1.0). Zenodo. doi:10.5281/zenodo.3727279

Environment Climate Change Canada (2016). Recovery Strategy for the Canada Warbler (Cardellina canadensis) in Canada. Species at Risk Act Recovery Strategy Series. Environment Canada, Ottawa. vii + 56 pp. Available at: http://www.registrelep-sararegistry.gc.ca, accessed February 13, 2020.

Fewster, R. M., S. T.Buckland, G. M.Siriwardena, S. R.Baillie, and J. D.Wilson (2000). Analysis of population trends for farmland birds using generalized additive models. Ecology 81:1970–1984. doi: 10.2307/177286

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. Bayesian Analysis. 1:515-533. doi:10.1214/06-BA117A

Gelman, A., J. Hwang, and A. Vehtari (2014). Understanding predictive information criteria for Bayesian models. Statistics and Computing 24:997-1016. doi:10.1007/s11222-013-9416-2

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari and D. B. Rubin. (2013). Bayesian Data Analysis , Chapman and Hall/CRC Boca Raton.

IUCN Standards and Petitions Committee. (2019). Guidelines for Using the IUCN Red List Categories and Criteria. Version 14. Prepared by the Standards and Petitions Committee. Downloadable from http://www.iucnredlist.org/documents/RedListGuidelines.pdf.

Lange, K. L., Little, R. J. A, and J. M. G. Taylor (1989). Robust Statistical Modeling Using the *t* Distribution. Journal of the American Statistical Association, 84:408, 881-896, DOI:10.2307/2290063

Knape, J. (2016). Decomposing trends in Swedish bird populations using generalized additive mixed models. Journal of Applied Ecology, 53, 1852–1861. doi: 10.1111/1365-2664.12720

Link, W., & Sauer, J. (2007). Seasonal Components of Avian Population Change: Joint Analysis of Two Large-Scale Monitoring Programs. *Ecology*, *88*(1), 49-55.

Link, W. A. and J. R. Sauer (2016). Bayesian cross-validation for model evaluation and selection, with application to the North American Breeding Bird Survey. Ecology 97:1746–1758. doi: 10.1890/15-1286.1

Link, W.A., J.R. Sauer, and D.K. Niven. (2017). Model selection for the North American Breeding Bird Survey: A comparison of methods. Condor 119(3):546–556. doi: 10.1650/CONDOR-17-1.1

Link, W. A., J.R. Sauer, and D.K. Niven. (2020 unpublished). Model selection for the North American Breeding Bird Survey, with observations on BPIC and WAIC model selection criteria.

Microsoft and Steve Weston (2019). foreach: Provides Foreach Looping Construct. R package version 1.4.7. https://CRAN.R-project.org/package=foreach

North American Bird Conservation Initiative Canada. (2019). The State of Canada's Birds, 2019. Environment and Climate Change Canada, Ottawa, Canada. 12 pages. www.stateofcanadasbirds.org

Partners in Flight. (2019). Avian Conservation Assessment Database, version 2019. Available at http://pif.birdconservancy.org/ACAD

Plummer, Martyn. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.

Pedersen EJ, Miller DL, Simpson GL, Ross N. (2019). Hierarchical generalized additive models in ecology: an introduction with mgcv. PeerJ 7:e6876. doi: 10.7717/peerj.6876

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Roberts, D. R., V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schroder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, and C. F. Dormann. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical or phylogenetic structure. - Ecography doi: 10.1111/ecog.02881.

Robbins, C. S., D. Bystrak, and P. H. Geissler (1986). The Breeding Bird Survey: Its first fifteen years, 1965–1979. U.S. Fish and Wildlife Service Resource Publication 157.

Rosenberg, K. V., P. J. Blancher, J. C. Stanton, and A. O. Panjabi (2017). Use of North American Breeding Bird Survey Data in avian conservation assessments. The Condor: Ornithological Applications 119:594–606. doi: 10.1650/CONDOR-17-57.1

Rosenberg, K. V., Dokter, A. M., Blancher, P. J., Sauer, J. R., Smith, A. C., Smith, P. A., Stanton, J.C., Panjabi, A., Helft, L., Parr, M., Marra, P.P. (2019). Decline of the North American avifauna. Science 366, 120–124. doi: 10.1126/science.aaw1313

Sauer, J. R., and W. A. Link. (2011). Analysis of the North American Breeding Bird Survey using hierarchical models. The Auk 128:87–98. doi: 10.1525/auk.2010.09220

Sauer, J.R., J.E. Hines, J.E. Fallon, K.L. Pardieck, D.J. Ziolkowski Jr., and W.A. Link. (2014). The North American Breeding Bird Survey, Results and Analysis 1966 – 2013. Version 01.30.2015 USGS Patuxent Wildlife Research Center, Laurel, MD.

Smith A.C., M.-A.R. Hudson, C. Downes, and C.M. Francis (2014). Estimating breeding bird survey trends and annual indices for Canada: how do the new hierarchical Bayesian

estimates differ from previous estimates. Canadian Field-Naturalist 128:119-134. doi: 10.22621/cfn.v128i2.1565

Smith, A. C., M.-A.R. Hudson, C. Downes, and C.M. Francis (2015). Change points in the population trends of aerial-insectivorous birds in North America: synchronized in time across species and regions. PLoS One 10:e0130768. doi: 10.1371/journal.pone.0130768

Smith, A.C., Hudson, M-A.R. Aponte, V., and Francis, C.M. (2019). North American Breeding Bird Survey - Canadian Trends Website, Data-version 2017. Environment and Climate Change Canada, Gatineau, Quebec, K1A 0H3

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*. 27(5), 1413--1432. doi:10.1007/s11222-016-9696-4.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York

Wilson, S., Smith, A. C., & Naujokaitis□Lewis, I. (2018). Opposing responses to drought shape spatial population dynamics of declining grassland birds. *Diversity and Distributions*, 24, 1687– 1698. doi: 10.1111/ddi.12811

Wood, S. N. (2017). Generalized additive models: an introduction with R; 2nd ed. CRC Press. Portland, OR, 2017

FIGURES

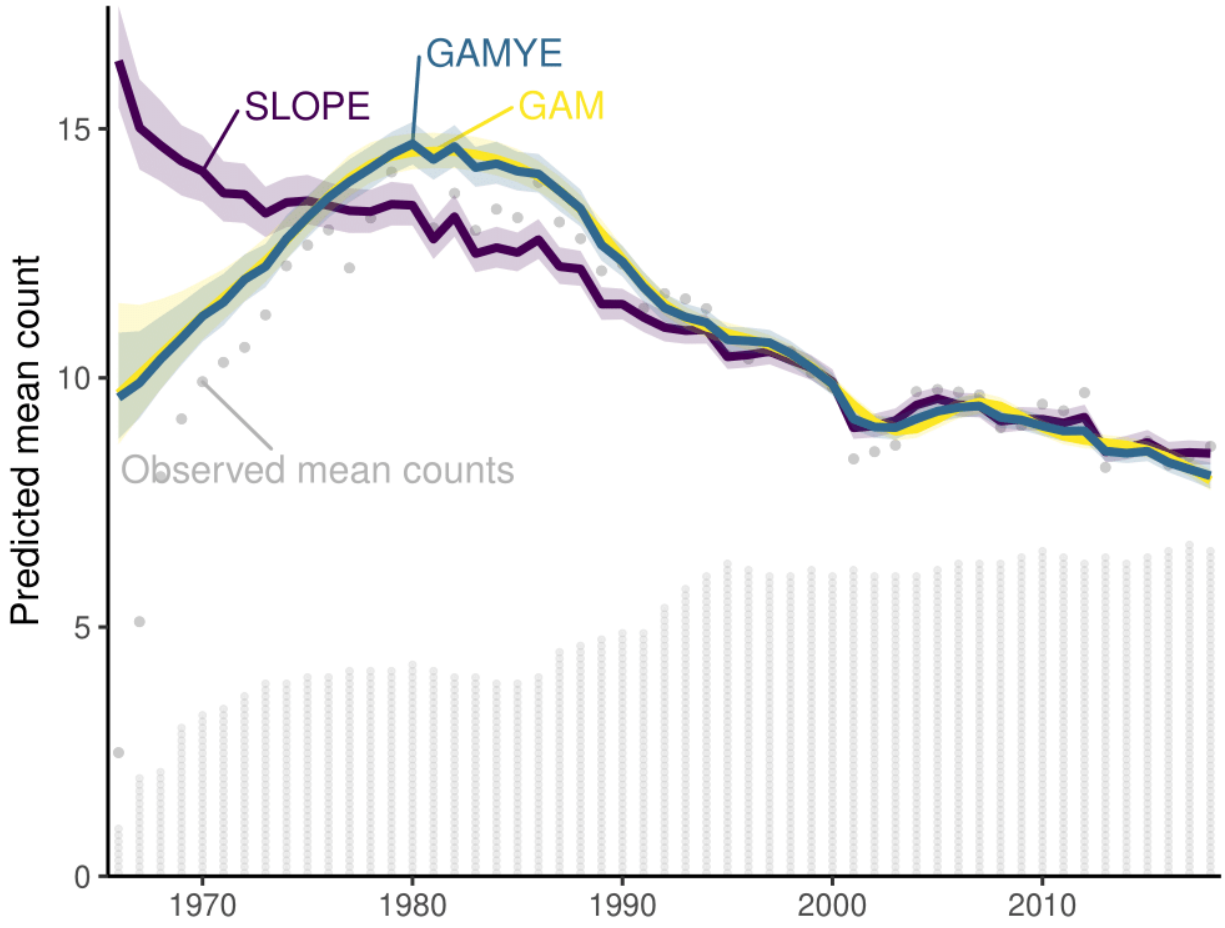

Figure 1. Survey-wide population trajectories for Barn Swallow (*Hirundo rustica*) estimated from the BBS using two models described here that include a GAM smoothing function to model change over time (GAM and GAMYE) and a third trajectory estimated using the standard slope-based model used for BBS status and trend assessments since 2011 (SLOPE). The stacked dots along the x-axis indicate the approximate number of BBS counts used in the model; each dot represents 50 counts.

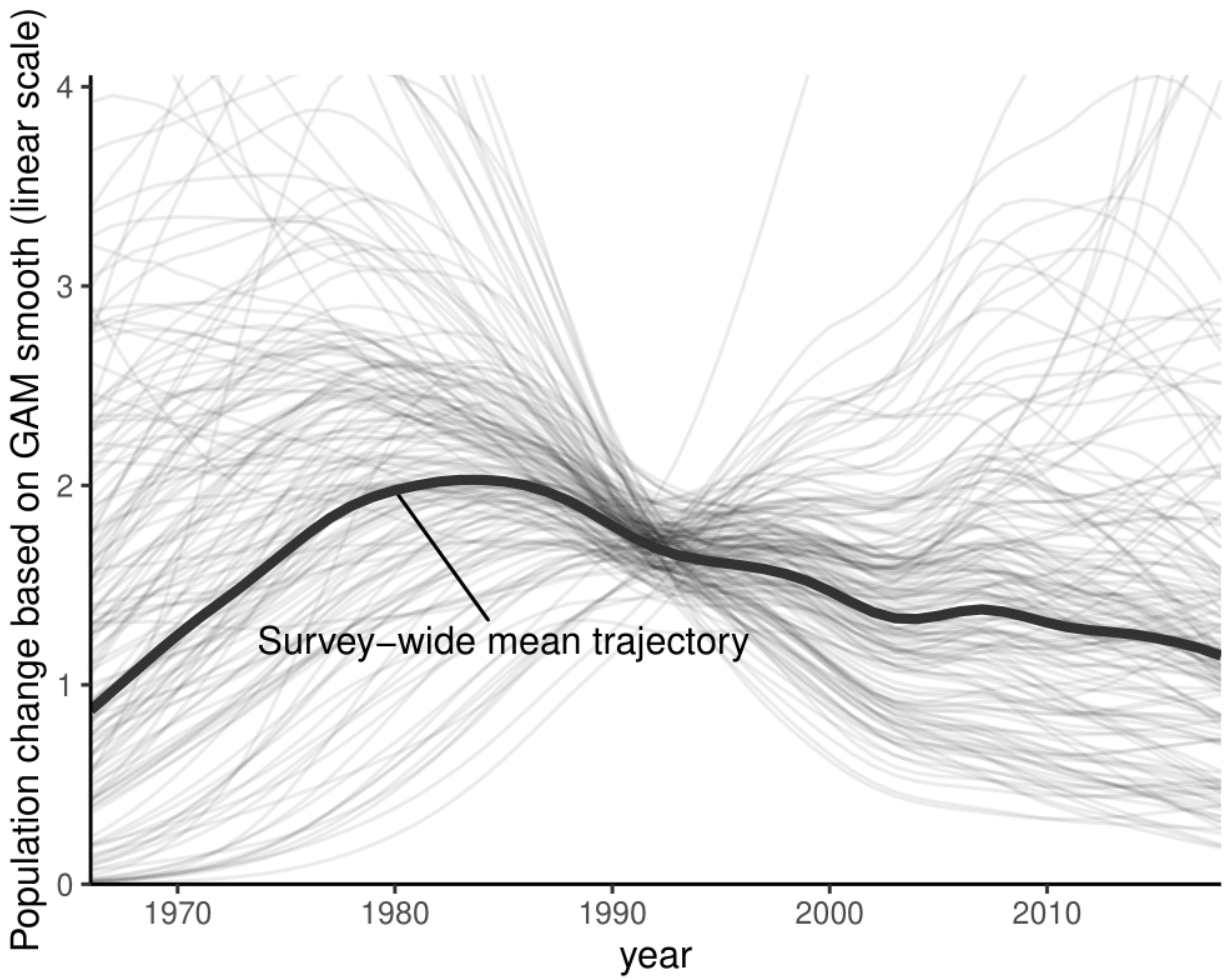

Figure 2. Variation among the spatial strata in the random-effect smooth components of the GAMYE model applied to Barn Swallow data from the BBS. Grey lines show the strata-level random-effect smooths, and the black lines shows the survey-wide mean trajectory.

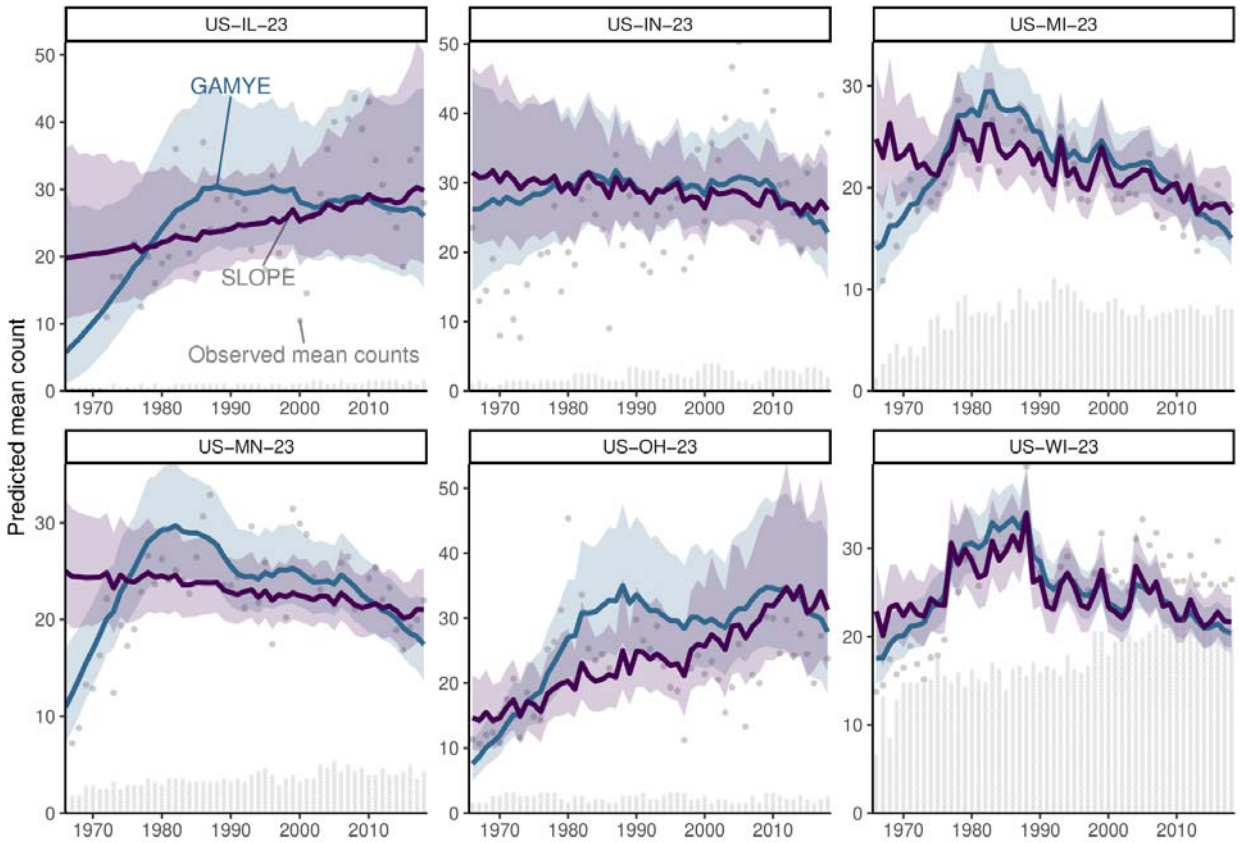

Figure 3. Stratum-level predictions for Barn Swallow population trajectories in BCR 23 from GAM and GAMYE against the predictions from the SLOPE model. The similarity of the overall patterns in the GAMs as compared to the SLOPE estimates, demonstrates the inferential benefits of the sharing of information among regions on the shape of the trajectory. In most strata the pattern of observed mean counts suggests a steep increase in Barn Swallows across all of BCR 23 during the first 10-years of the survey. The GAM and GAMYE estimates show this steep increase in almost all of the strata, whereas the SLOPE predictions only show this pattern in the most data rich stratum, US-WI-23. The facet strip labels indicate the country and state-level division of BCR 23 that makes up each stratum. The first two letters indicate all strata are within the United States, and the second two letters indicate the state. The stacked dots along the x-axis indicate the number of BBS counts in each year and stratum; each dot represents one count.

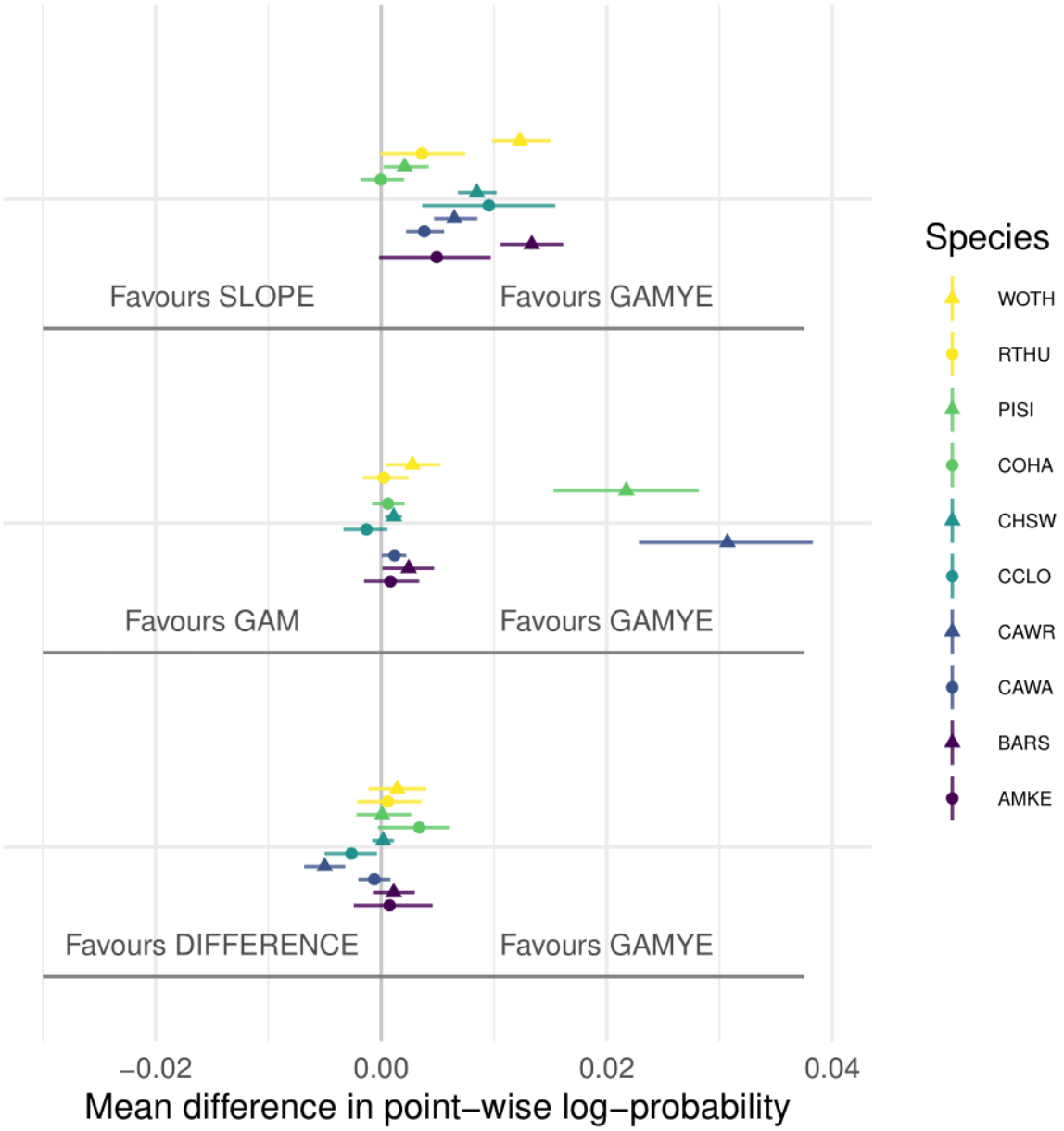

Figure 4. Overall differences in predictive fit between the GAMYE and SLOPE and GAMYE and GAM for Barn Swallow and 9 other selected species. Species short forms are WOTH = Wood Thrush (*Hylocichla mustelina*), RTHU = Ruby-throated Hummingbird (*Archilochus colubris*), PISI = Pine Siskin (*Spinus pinus*), Cooper's Hawk (*Accipiter cooperii*), CHSW = Chimney Swift (*Chaetura pelagica*), CCLO = Chestnut-collared Longspur (*Calcarius ornatus*),

CAWR = Carolina Wren (*Thryothorus ludovicianus*), CAWA = Canada Warbler (*Cardellina canadensis*), MAKE = American Kestrel (*Falco sparverius*).

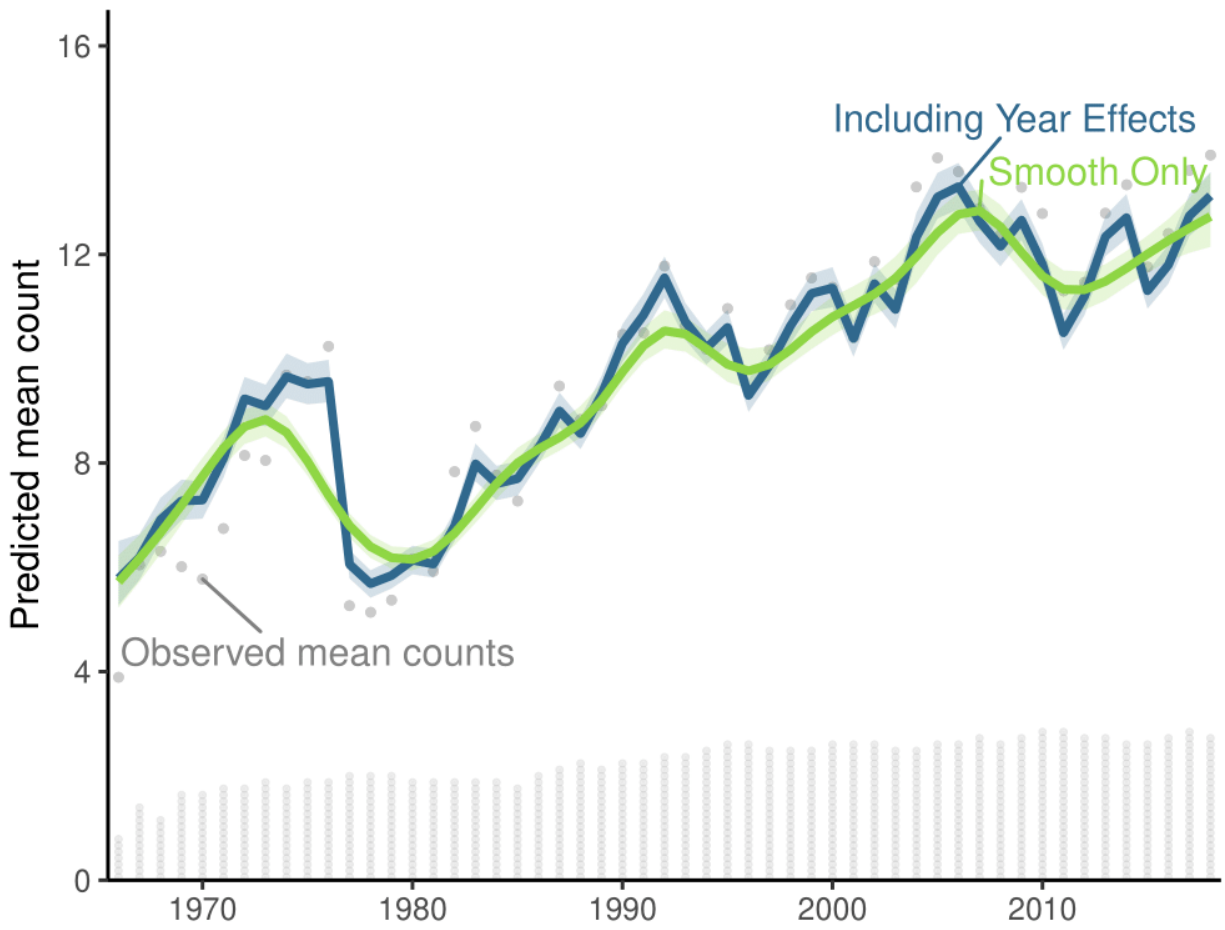

Figure 5. Decomposition of the survey-wide population trajectory for Carolina Wren (*Thryothorus ludovicianus*), from the GAMYE, showing the full trajectory ("Including Year Effects", $N_{s,t}$) and the isolated smooth component ("Smooth Only", $Ng_{s,t}$), which can be used to estimate population trends that are less sensitive to the particular year in which they are estimated. The stacked dots along the x-axis indicate the approximate number of BBS counts used in the model; each dot represents 50 counts.

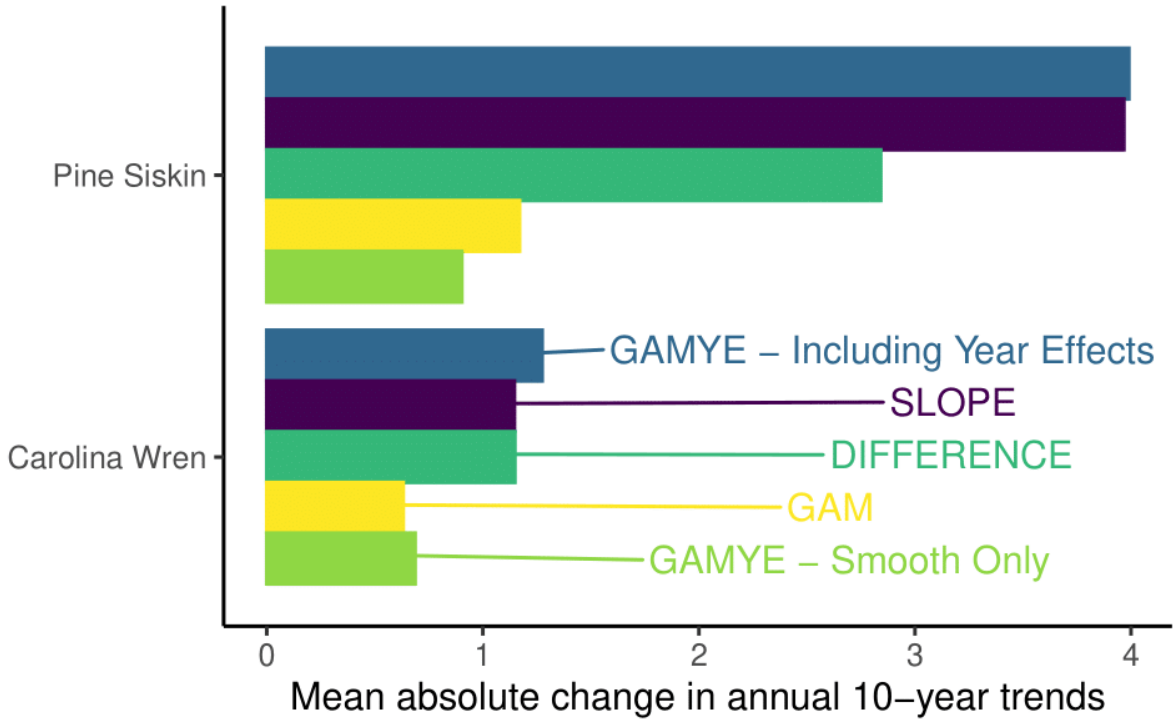

Figure 6. Inter annual variability of 10-year trend estimates for two species with large annual fluctuations. Trends from the GAM, which does not model annual fluctuations, and from the GAMYE using only the smooth component, which removes the effect of the annual fluctuations, are less variable between subsequent years (i.e., more stable) than trends from the GAMYE including the year-effects or the other two models that include the annual fluctuations.

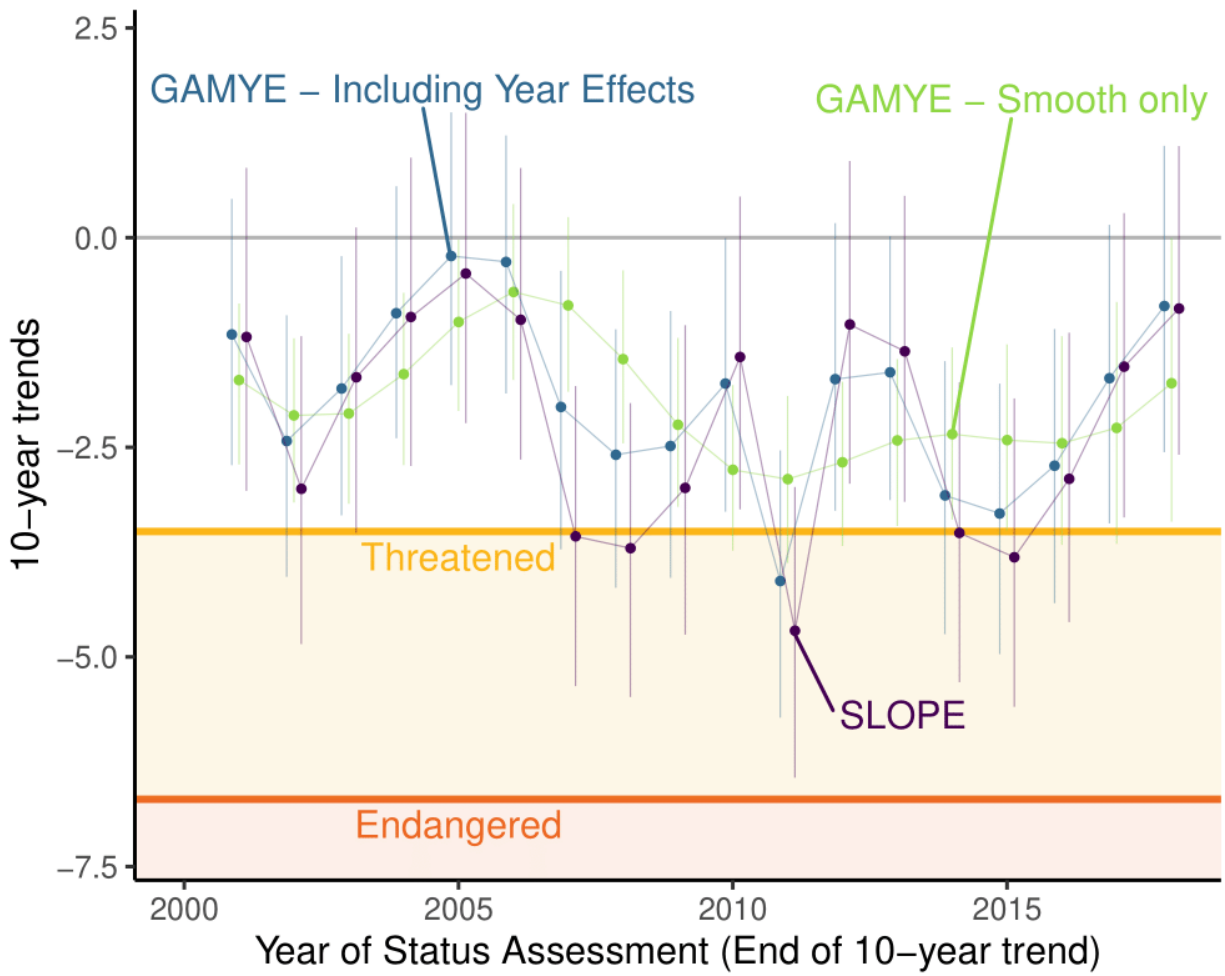

Figure 7. Sequential, short-term trend estimates for Wood Thrush (*Hylocichla mustelina*) in Canada from three alternative modeling approaches, and their comparison to the IUCN trend criteria for "Threatened" (in orange) and "Endangered" (in Red). Trends estimated from the decomposed trajectory of the GAMYE that include only the smooth component (in blue) are more stable between sequential yearly estimates than trends from either the GAMYE that include the year-effects or the SLOPE model that has been used to estimate BBS trends since 2011.

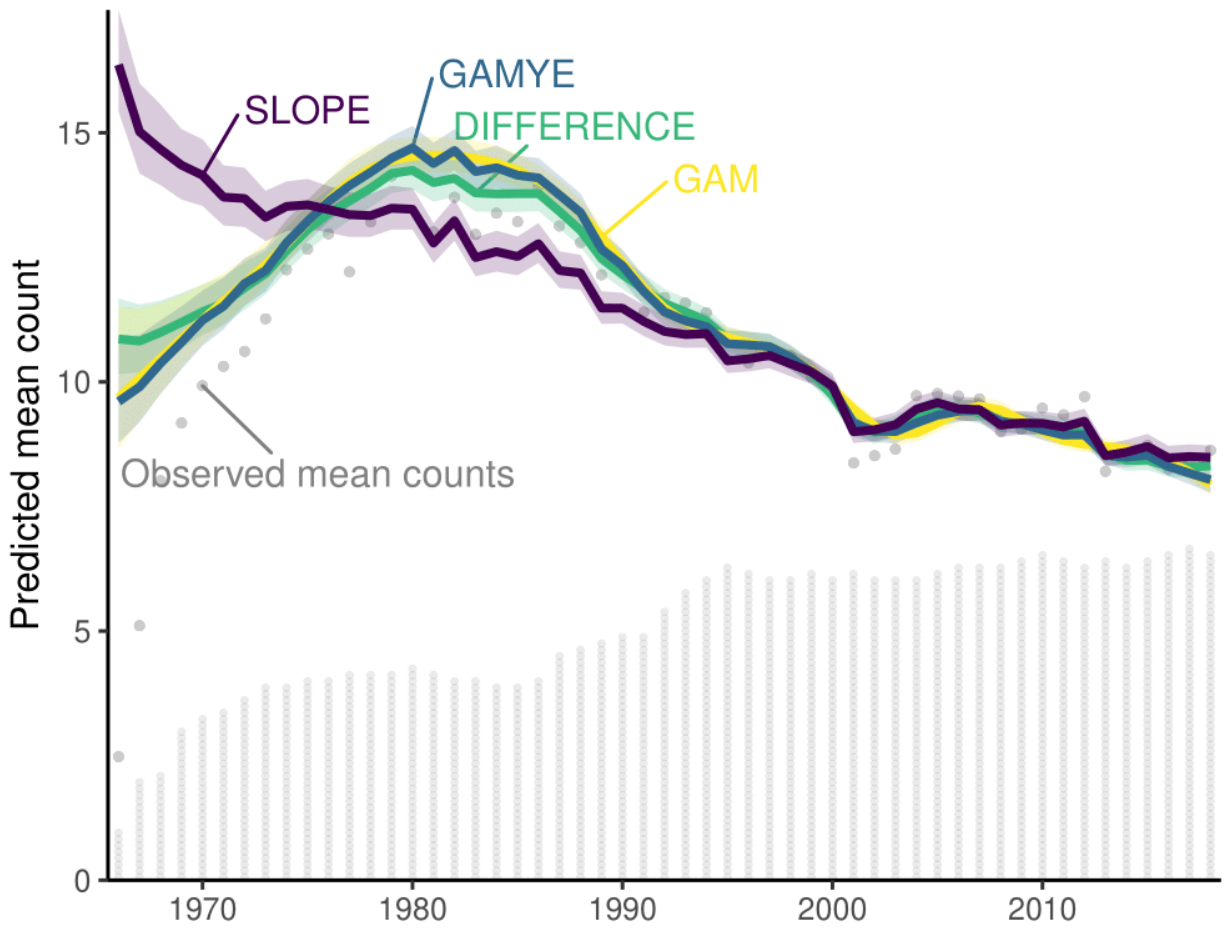

Figure 8. Predicted survey-wide population trajectories from four models applied to the Barn Swallow data from the BBS. The stacked dots along the x-axis indicate the approximate number of BBS counts used in the model; each dot represents 50 counts.

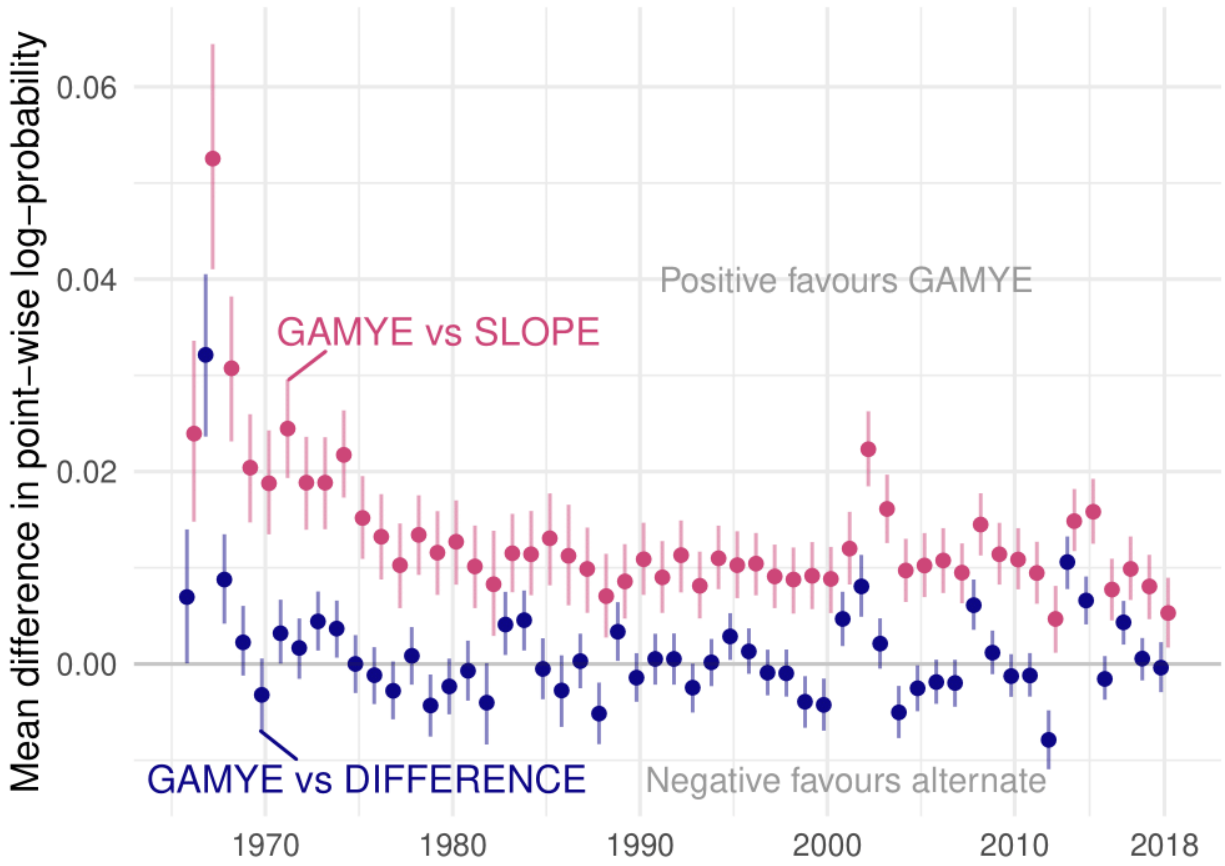

Figure 9. Annual differences in predictive fit between the GAMYE and SLOPE (blue) and the GAMYE and DIFFERENCE model (red) for Barn Swallow.

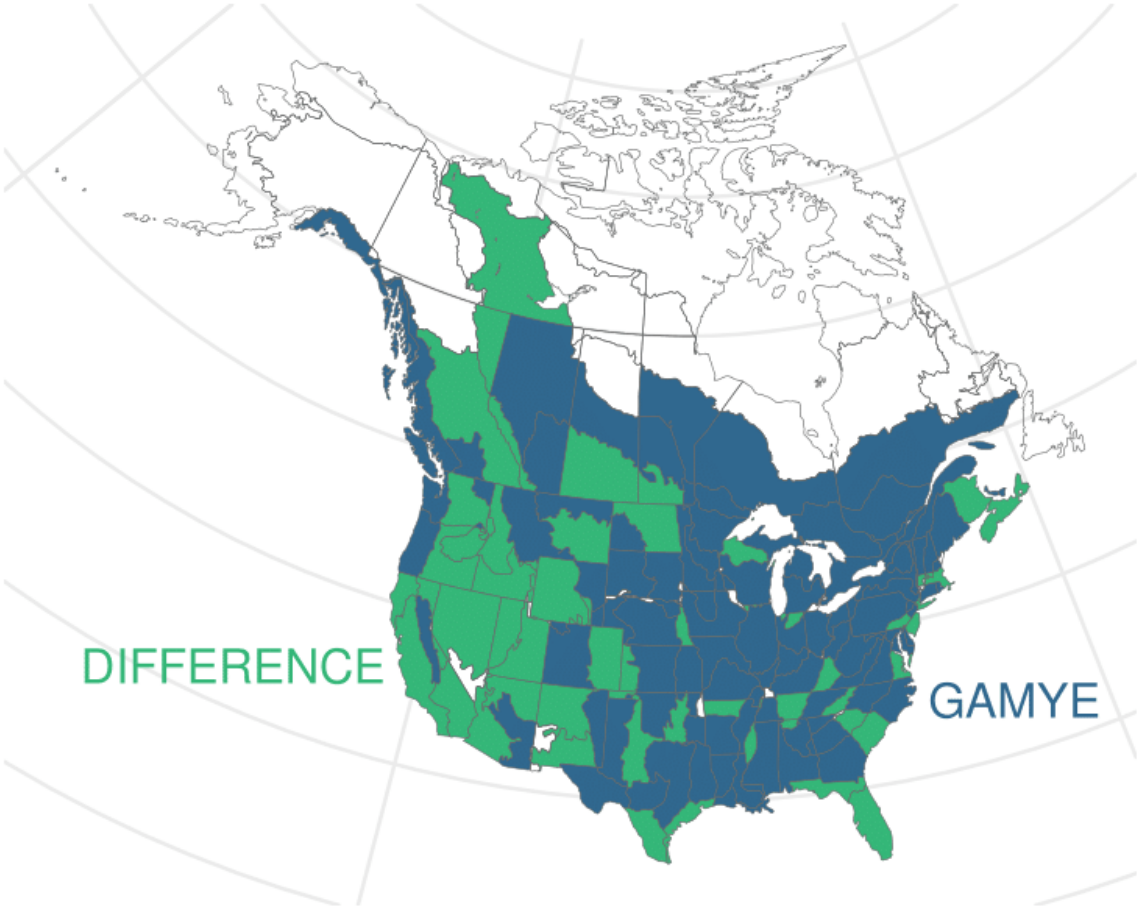

Figure 10. Geographic distribution of the preferred model for Barn Swallow, according to the point-estimate of the mean difference in predictive fit between GAMYE and DIFFERENCE. The GAMYE is clearly preferred in the Eastern part of the species' range, but the DIFFERENCE is preferred in many regions in the Western part of the species' range.