The Auk

Observer aging and long-term avian survey data quality

SCHOLARONE™
Manuscripts

1 **LRH**: FARMER ET AL.

2 **RRH**: OBSERVER AGING AND SURVEY DATA QUALITY

3 **Title:** OBSERVER AGING AND LONG-TERM AVIAN SURVEY DATA

4 QUALITY

5 **Authors:** R. G. Farmer[a], M. L. Leonard[a], J. E. Mills Flemming[b] and S. C. Anderson[a1]

6 [a]*Department of Biology, Dalhousie University, 1355 Oxford Street, Halifax, Nova Scotia,*

7 *Canada B3H 4R2*

8 [b]*Department of Mathematics and Statistics, Dalhousie University, 1355 Oxford Street,*

9 *Halifax, Nova Scotia, Canada B3H 4R2*

10 **Corresponding Author:** Robert G. Farmer

11 **Corresponding Address:** *Department of Biology, Dalhousie University 1355 Oxford*

12 *Street, Halifax, NS B3H 4R2*

13 **Corresponding Tel:** 902-494-2158

14 **Corresponding Fax:** 902-494-3736

15 **Corresponding email:** farmerb@dal.ca

---

[1]Current address: *Simon Fraser University, Dept. of Biological Sciences, Burnaby BC V5A 1S6*

Abstract

Long-term wildlife monitoring often involves volunteer observers collecting repeated, annual observations. However, when a volunteer participates for many years, uncorrected, age-related changes to his or her physiology might add negative biases to apparent population trajectories. Here, we used independent bird survey data from each of the Atlas of the Breeding Birds of Ontario ('OBBA') and the North American Breeding Bird Survey ('BBS') to show systematic declines in detection probabilities (OBBA) and expected counts (BBS) with increasing observer age. In a case study, we then showed how a failure to account for the continuous effects of observer age in a model of Golden-crowned Kinglet (*Regulus satrapa*) BBS count data led to more-negative population trajectory estimates. We tested for the importance of age-related hearing loss, which tends to affect higher frequencies at earlier ages, by asking if our observed detection declines were greater for species with higher call frequencies, and found some evidence of this effect for species with peak vocalization frequencies above 6 kHz. Among these same species, we also found that previously-published, long-term Canadian bird population trend estimates became increasingly negative as vocalization frequencies increased above 6 kHz, and as our corresponding estimates of observer-age-related detectability declines increased. Taken together, our results suggest that observer aging effects are important influences on long-term survey data quality, and that the mechanisms underlying this process include both age-related hearing loss and other physiological changes. Where possible, we recommend that survey designers and modelers account for observer age in future work.

KEY WORDS: avian ecology, bias, hearing loss, observer error, point counts, surveys

38   With support from thousands of volunteer observers, a variety of ongoing, long-term

39 ecological monitoring projects have dramatically improved our ability to address important

40 scientific questions (Silvertown 2009; Peters 2010). With data that can span decades and

41 continents (Silvertown 2009), these projects have become important components of wildlife

42 population assessment and management. For instance, data from the annual North American

43 Breeding Bird Survey ('BBS' [1966-present]; Peterjohn 1994) are used as part of avian

44 species-at-risk assessments (Greenberg and Droege 1999; Dunn 2002) and have also helped

45 to characterize the broad-scale effects of introduced species (Cooper et al. 2007), diseases

46 (LaDeau et al. 2007) and climatic variation (Link and Sauer 2007; Link et al. 2008; Wilson

47 et al. 2011). Similar projects have also helped to describe the distributions of anurans

48 (Blaustein et al. 1994; Lotz and Allen 2007), invertebrates (Kremen et al. 2011), and marine

49 life (Goffredo et al. 2010), among others. As with any scientific study, however, the quality

50 of the resulting inferences depends on the project's ability to control for errors that mask

51 true biological patterns.

52   Missed detections, which occur when an animal is present but not detected, are one such

53 class of error affecting monitoring projects (MacKenzie et al. 2005; Royle and Link 2006) and

54 can be caused by many environmental (e.g. Griffith et al. 2010) and observer-specific (e.g.

55 Sauer et al. 1994; McLaren and Cadman 1999; Alldredge et al. 2007b) factors. With present

56 statistical techniques, survey designs that involve repeated sampling, and supplementary

57 data collection, modelers can estimate and correct for missed detection rates directly (Royle

58 and Link 2006; Nichols et al. 2009; Miller et al. 2011, but see Campbell and Francis 2011).

59 Unfortunately, many point count surveys – including most BBS routes – lack this ideal design,

60 and so model-based covariates are often used to account for important sources of variation in

61 missed detection rates (e.g. Link and Sauer 2002). With this approach, if all sources of error

62 are recognized as covariates, the 'corrected' counts can then be used as unbiased indices of

63 overall population size (Johnson 2008).

64   In practise, however, some of these important covariates might be missed. For instance,

3

65 the current approach to analyzing BBS data taken by the United States Geological Service

66 (e.g. Link and Sauer 2002; Sauer and Link 2011) does not explicitly account for changes in

67 detection ability within particular observers over time (except for a first-year 'start-up' effect;

68 Kendall et al. 1996; Link and Sauer 2002). Hearing loss is one important physiological reality

69 that changes an observer's ability to detect sounds over time (Gates and Mills 2005), and

70 BBS data consist predominantly of aural detections (Cyr 1981; Faanes and Bystrak 1981).

71 Therefore, if declines in hearing ability as participants age are an important source of missed

72 detection errors, BBS (and similar) population models in their present form could in fact be

73 biased (sensu Simons et al. 2007).

74 Age-related hearing loss normally begins after age 30 for both men and women, and

75 includes a broadening range of frequencies beginning above 8 kHz, and progressing to frequencies

76 as low as 1 kHz by age 70 (Fig. 1; Mayfield 1966; Pearson et al. 1995). A second, less-common

77 form of hearing loss involves frequencies in a 'notched' range from 3 to 6 kHz (Nondahl et al.

78 2009; Osei-Lah and Yeoh 2010), which tends to appear more frequently with age (Toppila

79 et al. 2001). This 'notched' hearing loss is often associated with noise exposure, and its

80 increased prevalence among older people is likely a function of the cumulative effects of noise

81 and other stressors. Both forms of hearing loss are permanent (Gates et al. 2000; Wiley

82 et al. 2008; Cruickshanks et al. 2010), and they encompass the frequencies produced in many

83 bird vocalizations (Mayfield 1966; Emlen and DeJong 1992), including a group of warblers,

84 nuthatches and flycatchers which we studied here (Appendix 1).

85 In the BBS, which is a large and influential source of North American bird population

86 data (Sauer and Link 2011), the majority of observers are over 45 years of age (Fig. A1 in

87 Appendix 2; and see Wiedner and Kerlinger 1990; La Rouche 2001; Downes 2004; Carver

88 2009), and they tend to survey routes repeatedly for many years. During this time, aging-related

89 physiological changes inevitably take place. Given the age-associated changes in human

90 hearing, we hypothesize that this aging population of BBS volunteers may detect fewer birds

91 because of an inability to hear them, independent of real ecological patterns. If so, observed

92 population declines derived from these data may in fact reflect reduced detection abilities,

93 independent of any real biological change. With thousands of volunteers playing important

94 roles in population monitoring in the BBS and similar studies (e.g. the Audubon Christmas

95 Bird Count; La Sorte and McKinney 2007), this is a potentially serious issue.

96 Several studies have already argued that changes in observer hearing ability might bias

97 current models of bird species counts (e.g. Faanes and Bystrak 1981; Ramsey and Scott 1981;

98 Emlen and DeJong 1992; Simons et al. 2007), although these have largely been discussion

99 points or very limited in their quantitative scope. For instance, Faanes and Bystrak (1981)

100 considered 15 years' of data comparing relative counts among 3 BBS observers (routes),

101 showing negative differences in expected counts between hearing-impaired ($n = 1$) and

102 unimpaired ($n = 2$) observers. Similarly, Ramsey and Scott (1981) measured widespread and

103 variable hearing deficits in a large selection of active birders ($n = 274$), but only considered

104 the functional impact of this phenomenon on population trend estimates using theoretical

105 projections. Emlen and DeJong (1992) observed hearing differences among 2 observers aged

106 25 and 70, but again only speculated on its impact in ornithology using theoretical measures.

107 Using a controlled field set-up, Simons et al. (2007) showed substantial underestimation

108 of abundance of the high-frequency Black-and-white Warbler (*Mniotilta varia*) among 15

109 observers, and argued that age-related hearing loss is probably driving this pattern, but

110 again did not explicitly measure observer age and the associated age-dependent detection

111 patterns.

112 To our knowledge, only Link and Sauer (1998) considered observer age and its functional

113 impacts explicitly using a dataset containing multiple observers. In this case, the authors

114 predicted a "43% diminution of counts" for Blue-gray Gnatcatchers (*Polioptila caerulea*)

115 among BBS observers who have conducted surveys for more than 20 years. This "observer

116 senescence effect" – which here was a very-limited discussion point in a much-broader research

117 paper – has not been further explored in the published literature. Hence, our understanding

118 of the biases resulting from hearing loss and other aging phenomena in models of long-term

5

119 survey data is still quite poor, in spite of the potential for such errors to affect population

120 trend estimates and management actions. Furthermore, any such problem is unlikely to

121 disappear, as it is intrinsic to any survey route visited exclusively and repeatedly by a single

122 observer, who must necessarily age over the course of such service. We hence believe that there

123 is a critical need for future research that explicitly measures any detection biases resulting

124 from observer aging.

125 Our goals for this study were to test for the existence and consequences of age-related

126 declines in the detection abilities of long-term bird survey observers, with a focus on hearing

127 loss as a potential mechanism. We used data from two independent volunteer bird survey

128 datasets: the Atlas of the Breeding Birds of Ontario ('OBBA'; Bird Studies Canada et al.

129 2008) and the North American Breeding Bird Survey (Peterjohn 1994) to establish how bird

130 detection probabilities and expected species counts, respectively, tend to change as observers

131 age. We considered the role of hearing loss by simultaneously testing for patterns between

132 age-related differences in detection ability and species vocalization frequencies. We expected

133 to see the strongest effects for species with vocalization frequencies corresponding to common

134 age-associated hearing impairments.

135 We then asked how age-related changes in observer ability might bias long-term estimates

136 of population change. We first conducted a case study using BBS counts of the Golden-crowned

137 Kinglet ('GCKI'; *Regulus satrapa*), a small songbird chosen because of its high-frequency

138 vocalization (Appendix 1) and relatively high mean abundance. Here, we illustrated how

139 using uncorrected data from aging observers might bias population trajectory estimates.

140 We then considered whether hearing loss effects might be adding error to existing

141 long-term population trend estimates. In this case, we tested for a relationship between

142 previously-published, long-term population trends and the vocalization frequencies of each

143 corresponding species. We expected that population trend estimates would be more negative

144 in species with vocalization frequencies corresponding to common forms of hearing loss.

145 To measure the effects of generalized observer aging on population trend estimates, we

6

146 lastly tested whether our estimates of age-related differences in detection among younger

147 and older observers (derived from the OBBA data) might be correlated with long-term

148 population trend estimates (derived from BBS data). Here, we expected that population

149 trends estimates would be more negative for species showing greater declines in detectability

150 for older observers.

METHODS

151

Calculating and Classifying Vocalization Frequencies

152

153 In all analyses, we focused on a group of North American songbirds (warblers, nuthatches,

154 flycatchers) for which we could obtain high-quality vocalization data (Appendix 1). These

155 species form a major proportion of North American breeding bird species (e.g. `http://`

156 `www.aba.org/checklist/`), they represent a broad range of vocalization frequencies (Brand

157 1938), and they are frequently of conservation interest (e.g. Faaborg et al. 2010).

158 We categorized these species according to their acoustic characteristics by first determining

159 the peak (i.e. dominant) frequencies for the vocalizations of each species following Emlen and

160 DeJong (1992, see Appendix 3 for details). We then used these peak frequency values to assign

161 species into one of four vocalization frequency groups corresponding to known hearing-loss

162 thresholds of 3 kHz, 6 kHz and 7 kHz (International Organization for Standardization

163 2000). Accordingly, species were considered to have 'low' (< 3 kHz), 'notch' (≥ 3 and

164 < 6 kHz), 'medium' (≥ 6 and < 7 kHz) and 'high' (≥ 7 kHz) vocalizations. We hypothesized

165 that the detection of 'high' vocalizations (≥ 7 kHz) is most likely to change with age (a

166 result of age-related hearing loss; e.g. Fig. 1; International Organization for Standardization

167 2000; Gates and Mills 2005); whereas 'notch' vocalizations (3 to 6 kHz) may also show

168 age-associated detection declines (cumulative noise-induced and other idiopathic hearing

169 losses; Nondahl et al. 2009; Osei-Lah and Yeoh 2010). 'Medium' and 'low' vocalizations

170 lie between the 'high' and 'notch' categories.

171 'Peak frequency' is most representative of a particular vocalization if the vocalization

172 broadcasts a very narrow range of frequencies overall (Fig. A2 in Appendix 2; Ramsey and

173 Scott 1981). By comparison, this single statistic is much less relevant to vocalizations

174 incorporating several high-energy sounds at many disparate frequencies. To account for

175 this difference and focus on the former type, we thus classified vocalizations as being either

176 'monotone' or 'heterogeneous' according to the range of frequencies found in each power

177 spectrum (standard deviation of power values). We defined 'monotone' vocalizations as

178 those vocalizations with power spectra having a standard deviation less than or equal to

179 the median value among a group of 94 species initially considered; all other vocalizations

180 having more variable power spectra were defined as being acoustically 'heterogeneous'. Using

181 these heterogeneity classes, we thus expanded the existing four vocalization groups discussed

182 above into eight (i.e. 'High Monotone', 'High Heterogeneous', 'Medium Monotone', 'Medium

183 Heterogeneous', 'Notch Monotone', 'Notch Heterogeneous, 'Low Monotone', 'Low Heterogeneous';

184 Appendix 1). We expected that any relationships between species detections and vocalization

185 frequencies resulting from frequency-specific hearing loss phenomena would be stronger among

186 monotone species.

187 This simple classification method did not recognize cases where bird vocalizations

188 featured a wide range of frequencies broadcast over a very short time interval (e.g. Least

189 Flycatcher [*Empidonax minimus*]) – and so which appear subjectively monotone to the human

190 ear in spite of their having a heterogeneous power spectrum. However, this error did not risk

191 the inclusion of subjectively heterogeneous species in the monotone groups – a more serious

192 error because we were largely concerned with patterns among monotone species only – and

193 so this error was a conservative one.

Determining Observer-Age-Related Differences in Species Detections

195 *Age-related differences in OBBA detection probabilities.* — To estimate the difference

196 in detection probability between older and younger observers (defined below), we used data

8

197 from 43 species surveyed as part of the OBBA which had at least 100 detection records in total

198 (Appendix 1), and for which we were able to determine peak vocalization frequencies (see

199 above). The OBBA is a volunteer survey that divides the entire land area of the Canadian

200 province of Ontario into a grid of 3,324 10 km × 10 km squares, and during two 5-year

201 periods ('first atlas': 1981–1985; 'second atlas': 2001–2005), one to several volunteers per

202 square conducted area searches for breeding evidence of bird species during the spring and

203 summer months, with a minimum effort of 20 party-hours per square.

204   Working with atlas squares sampled during at least two separate years by one or more

205 observers, we inferred species detections as occurring when an atlasser reported any evidence

206 of a given species in a given atlas square. Conversely, we inferred nondetection for a given

207 atlasser and species by determining all squares visited by an atlasser, and assigning zeroes

208 ('no detection') to all species that were not reported there (sensu Kéry et al. 2010).

209   We used publicly-available data sources, including OBBA results web pages and field

210 naturalist groups' newsletter reports, to determine the approximate ages (under-40, 40–50, or

211 over-50) during the midpoint of the second atlas (2003) for 626 of 1,230 atlassers (demographic

212 data were not available for many observers). Although our primary interest was in measuring

213 age-related patterns of detection ability, we also recognized that gender could have an

214 influence, because men tend to lose their high-frequency hearing sooner than women (Fig. 1),

215 and so we recorded gender as well. We also corrected for observer effort, both by excluding

216 records with zero-values and by modeling species detectability with effort as a covariable (see

217 below).

218   We could not explicitly distinguish between casual, "backyard" observations and more-distant

219 field searches within a given atlas square, the former of which might be more often associated

220 with older, less-mobile birdwatchers. Any such relationship, if widespread in the data, could

221 confound age-related differences in observer sensory abilities (of interest here, and relevant

222 to fixed-protocol surveys like the BBS) with age-related differences in observer mobility.

223 However, we were unlikely to successfully determine an observer's age and hence, include

9

224 that observer's data, unless he or she had a substantial field naturalist group presence (for

225 instance, sufficient to warrant publishing his or her name and photograph in a newsletter),

226 and in our experience, active field naturalist group participation implies an ability and a

227 preference to visit sites further afield than a backyard. Hence, we suspect that this potential

228 confounding influence is not widespread in the data used here.

229 To model the effects of observer age on detection probability, we constructed hierarchical

230 occupancy models of the resulting dataset (Royle and Kéry 2007; Royle and Dorazio 2009)

231 in WinBUGS 1.4.3 (Lunn et al. 2009) and R 2.13.0 (R Development Core Team 2011)

232 using the R package *arm* (Gelman et al. 2010) on a PC running Windows 7. Hierarchical

233 occupancy models simultaneously estimate species presence ('occupancy') and detection

234 probability (conditional on occupancy), along with the effects of specified covariates, from

235 detection/nondetection datasets that have been repeatedly sampled at a set of locations (i.e.

236 atlas squares). Consistent with the assumptions of OBBA design, we assumed that occupancy

237 did not change for a given atlas square during each of the 5 years of an atlassing period, and

238 so treated each atlas square as a sampling unit and each sampling year as a within-observer

239 replicate. We did not account for the presence of any false-positive detections in the data

240 (e.g. McClintock et al. 2010*b*).

241 In the models, we used up to 350 observers (depending on the species being modeled)

242 from two age groups, observers under 40 (18 women, 65 men) and observers over 50 (64

243 women, 203 men), and expected that the older cohort would have functionally reduced

244 hearing compared to the younger one, on average (Fig. 1). We avoided using records for

245 atlassers we believed to be aged between 40 and 50 years in order to preserve such an expected

246 functional distinction (sensitivity analyses later validated this concern; see below), and to

247 account for any errors in age determination. Each hierarchical occupancy model consisted of

248 an occupancy component, which predicted true occupancy in the second atlas as a function of

249 detection in the first atlas, and a detection component, which predicted detection (conditional

250 upon occupancy) as a function of effort (survey hours), observer gender, observer age (over

₂₅₁ 50 vs. under 40), and random observer variation. Specific formulations of the occupancy

₂₅₂ models and Bayesian priors used are discussed in Appendix 4.

₂₅₃   In each converged occupancy model, the 'β_2' 'observer age' parameter (see equation 5

₂₅₄ in Appendix 4) corresponded to the difference in detection probability on the logistic scale

₂₅₅ between observers younger than 40 and older than 50 for a given species. The mean of all

₂₅₆ β_2 estimates describes how observer detection ability is expected to differ between older and

₂₅₇ younger observers among all species considered.

₂₅₈   We then described the role of hearing loss in driving age-related differences in this

₂₅₉ detection probability statistic, if any, by constructing an additive model ('GAM'; Wood 2006)

₂₆₀ using the R package *mgcv* (Wood 2006) to predict the 43 estimates of β_2 as a function of the

₂₆₁ peak vocalization frequencies of each corresponding bird species. Compared to parametric

₂₆₂ approaches, which require polynomial curve orders to be defined *a priori*, GAMs can fit

₂₆₃ nonparametric smooth functions predicting the most likely nonlinear relationships between

₂₆₄ x (i.e. vocalization frequency) and y (i.e. β_2). The package *mgcv* optimizes the amount

₂₆₅ of 'wiggliness' using internal cross-validation algorithms (Wood 2006). In these models,

₂₆₆ P-values correspond to the probability that a smooth function exists by random chance (i.e.

₂₆₇ vs. no linear nor nonlinear relationship in the data). The resulting smooth functions are

₂₆₈ intended to be visually inspected by the modeler; summary statistics alone are inadequate

₂₆₉ to describe their findings.

₂₇₀   We considered relationships between β_2 and vocalization frequency for the monotone

₂₇₁ and heterogeneous vocalizations separately, and weighted the datapoints according to the

₂₇₂ inverse of the variance of their posterior distributions (i.e. their uncertainty; estimated earlier

₂₇₃ by WinBUGS). If age-related hearing loss is an important mechanism leading to age-related

₂₇₄ detection declines, we expected the magnitude of the β_2 values to decline with increasing

₂₇₅ vocalization frequency. Declines in the 'notched' region would similarly correspond to an

₂₇₆ influence of noise-induced hearing loss. We also expected to see more-pronounced patterns

₂₇₇ for the 'monotone' species groups, because in this case, the peak vocalization frequency more

278 closely corresponds to the principal frequency heard by the observer.

279 To justify our exclusion of observers from the 40 to 50 age group, and furthermore, to

280 validate whether observer age influences detection ability, we also tested for the sensitivity of

281 the values of β_2 to the inclusion of observers of borderline age (i.e. ages 40–50). We re-fit the

282 occupancy models as described above, except here using data from observers of all ages (while

283 retaining the age-50 cutoff), and then compared matched pairs of these new β_2 estimates to

284 their earlier estimates. If there are important differences in detection ability that develop

285 progressively by age 50, we would expect to see a smaller overall difference in detection ability

286 between the under-50 and over-50 cohorts, compared to the differences previously measured

287 between the under-40 and over-50 cohorts.

288 *Expected BBS counts derived from BBS data.* — Next, to compare the patterns observed

289 above with an independent dataset, we determined how bird counts varied with increasing

290 observer age on the BBS. Like the OBBA, the BBS is a multi-year, omnibus bird survey.

291 Here, it is conducted by skilled volunteers annually during the breeding season. In contrast to

292 the OBBA, most BBS survey stops are not replicated within survey cycles or among multiple

293 observers. The BBS uses a set of permanent, 39.4 km road transects ('routes') which are

294 divided into 50 stops placed at regular ($\sim$ 800 m) intervals. Most BBS routes are sited

295 randomly within North American physiographic subregions ('strata'; e.g. 'Sierra Nevada';

296 'St. Lawrence River Plain') and within degree blocks of latitude and longitude (Sauer et al.

297 2003), and so have a nested random structure. Survey routes continue to be added to the

298 BBS as a whole; the oldest routes have been monitored annually since 1966.

299 In the raw BBS count dataset, observers are assigned a unique identification number

300 which persists throughout their years of service. We used these identification numbers to

301 determine a measure of "minimum observer age", defined as the number of years since the first

302 year an observer served on any BBS route (sensu Faanes and Bystrak 1981). Within observers,

303 minimum observer age is correlated with actual observer age – our latent variable-of-interest

304 – by definition. We recognize that this measure is less-precise than true age, however for

305 simplicity, we refer to 'minimum' observer ages, which range from 1 to 39 in the data, as
306 'observer ages'.

307   We omitted data from the early years of the BBS, and instead used data collected in
308 Canada and the USA between 1970 and 2007 by single observers under suitable weather
309 conditions. These omissions avoided potential problems with low observer quality in the
310 early years of the survey (e.g. Sauer et al. 1994), as well as problems with 'anomalous results'
311 with early data from Canadian survey routes (see `http://ec.gc.ca/reom-mbs/default.`
312 `asp?lang=En&n=E8974122-1`). Observer ages were calculated using the original, complete
313 dataset, which began in 1966.

314   To account for changes in each species' population abundance occurring alongside
315 changes in observers' detection ability, we needed replicated time series for each location.
316 To achieve this effect with the unreplicated BBS data, we estimated population trajectories
317 at the level of the physiographic stratum, using count data from the individual survey routes
318 as replicates. We required that at least 3 separate observers be associated with a given
319 stratum before that stratum was included in the analysis. Similarly, to ensure that the age
320 and population effects under study were not confounded, we required that for each species
321 and stratum analyzed, observer age and calendar year were correlated by no more than 0.7
322 (Pearson correlation). We also worked exclusively with contiguous observer-route time series
323 of 10 years or longer, both to minimize biases that could result from any gaps in temporal
324 coverage (e.g. Sauer et al. 1994), and to capture age-related changes in detection ability.

325   Because raw BBS data do not include zero-counts for any species, we added relevant
326 zeroes in the same manner as was done with the OBBA dataset (sensu Kéry et al. 2010). The
327 presence of zero values on a route-year time series did not affect whether it was considered
328 contiguous or not. To avoid problems with estimating zero-values on the (logarithmic) scale
329 of the linear predictor in these models, we also added a value of 0.5 to all counts (Sauer et al.
330 1996).

331   Volunteer BBS observers often perform relatively poorly during their first year on a

13

survey route compared to later years; this phenomenon can inflate population trend estimates

if the first year of data is included (Kendall et al. 1996). To avoid confounding this pattern

with hearing loss phenomena, we excluded the first years' datapoints (mean 6.1% of records

per species) for each observer-route combination. Final datasets for each of 65 species meeting

these requirements (and for which we had vocalization frequency data; Appendix 1) ranged in

size from 37 (Lucy's Warbler [*Oreothlypis luciae*]) to 6692 (Common Yellowthroat [*Geothlypis*

trichas]) route-years of data, with a median 1077 records (mean 1608 $\pm$ 1689 [SD]).

We used overdispersed Poisson generalized additive mixed models ('GAMMs'; Wood

2006) in R package *gamm4* (Wood 2011) to model the nonlinear changes to BBS counts

with observer age, while controlling for both among-observer effects and continuous changes

in species counts with calendar year within physiographic strata (i.e. 'population' changes).

GAMMs are extensions of GAMs which incorporate additional random-effects structures

to account for group-specific deviations from overall means ('random intercepts') and from

overall trends ('random slopes'). These models are suitable for the hierarchical structure of

BBS data. As with GAMs, the 'significance' of a smooth function alone can only establish

whether there is a nonzero pattern to the data – 'significant' GAMMs are not necessarily

unidirectional (i.e. exclusively increasing or decreasing). Hence, GAM and GAMM smooths

are visual instruments, and their shapes must always be examined in order to properly

interpret their findings.

We aggregated the species-specific GAMMs produced here by building a second GAMM

describing the overall proportional changes to modeled BBS counts with increasing observer

age (relative to each species' values at age-1) for groups of species of the same vocalization

frequency and heterogeneity groups defined earlier. Finer details of this modeling process

are outlined in Appendix 5.

Species Detection Probabilities and Long-term Population Trend Estimates

Golden-crowned Kinglet case study. — We used BBS counts of the Golden-crowned

358 Kinglet ('GCKI') as the subject of a case study to examine how failing to control for

359 observer age could affect estimated population trajectories. Our goal was to model population

360 trajectories in a manner consistent with established techniques (Link and Sauer 2002; Sauer

361 and Link 2011), and then to compare the resulting estimates made both with and without a

362 continuous correction for minimum observer age.

363   Here, we excluded continuous observer age effects from the GCKI GAMM produced

364 in the earlier BBS count analysis (see above), producing a model which approximated the

365 hierarchical Bayesian modeling methods presently used by the US Geological Service ('USGS';

366 Link and Sauer 2002). The only major difference between this second GAMM approach and

367 the USGS models is the GAMMs' use of a smooth function for calendar year in place of

368 year-specific parameters to describe annual changes. The numerous parameters in the USGS

369 approach should serve roughly the same purpose as a continuous GAMM function, and so

370 predictions from both our GAMMs (modeled without observer age corrections) and the USGS

371 hierarchical Bayesian models should be largely equivalent given the same initial dataset.

372   We compared the shapes of the resulting smooth functions for 'population' trajectories

373 ($f_2(l)_j$ in Equation 8 in Appendix 5) between GAMMs built with and without age corrections

374 for three physiographic strata that represent both apparently stable and apparently declining

375 'uncorrected' population trajectories (Northern Spruce-Hardwoods [stable], Sierra Nevada

376 [declining], South Pacific Rainforests [declining]). Because we used the BBS data subset

377 discussed earlier, our population estimates do not necessarily reflect true biological patterns

378 (e.g. Sauer et al. 1996), but rather they illustrate the directional effects of observer age

379 corrections on population trajectory inferences under severe circumstances (i.e. where most

380 modeled observers experience long-term aging).

381   *The influence of aging observers on population trend estimates.* — Finally, we determined

382 if aging-related observer effects might have influenced existing broad-scale population trend

383 estimates. Here, we considered Canada-wide population trend statistics produced by both the

384 USGS (http://www.mbr-pwrc.usgs.gov/cgi-bin/atlasa09.pl?CAN&2&09), and by the Canadian

385 Wildlife Service ('CWS '; http://www.cws-scf.ec.gc.ca/mgbc/trends/index.cfm?lang=

386 e&go=info.SpeciesListByProvince&provid=0). Both sets of trends are calculated by their

387 respective agencies using area-weighted, Poisson-modeled BBS count data, where estimated

388 'trend' values correspond to the estimated exponential rate of change of a population from the

389 beginning to the end of the survey period modeled. However, fine details of these strategies

390 are not equivalent. Thomas and Martin (1996) showed that agency-specific differences in

391 such analysis strategies (i.e. different geographic weighting schemes) can lead to important

392 differences in trend magnitude and significance. Current trend estimation strategies have

393 improved since 1996 among both agencies, but remain divergent for other reasons (C. Francis,

394 pers. comm.). Here, we wanted to determine if there was an agency-independent (i.e.

395 common) effect of vocalization frequency among each set of trends, and so considered both.

396   If hearing loss is an important predictive factor, we expected to see greater estimated

397 population declines among species with vocalization frequencies associated with hearing loss.

398 To test for such a pattern, we built single-parameter GAMs relating each of the USGS and

399 CWS population trends with species vocalization frequency, specifying separate thin-plate

400 regression spline smoothers for monotone and for heterogeneous vocalizations. The USGS

401 dataset supplied 95% credible intervals about the trend estimates; consequently, we treated

402 the width of these intervals as a measure of error and weighted datapoints according to their

403 inverses. Similarly, we used the supplied number of BBS routes incorporated into each CWS

404 trend prediction as a corresponding weight in the CWS GAM. We used population trend

405 estimates that spanned the longest available timespan in each case, which was from 1966 to

406 2009 for the USGS trends ($n = 50$ warbler, flycatcher and nuthatch species for which we had

407 vocalization frequency data), and from between 1970 and 1973 to 2009 for the CWS trends

408 ($n = 52$ species). We excluded one CWS trend (Bohemian Waxwing [*Bombycilla garrulus*])

409 which was valid only for 1986 to 2009.

410   More broadly, to determine if observer aging in general might be biasing population

411 trend estimates, we then compared our estimates of observer-age-related differences in detection

412 probability (i.e. β_2; OBBA data) to the USGS and CWS population trend estimate data

413 (derived largely from BBS data; Sauer and Link 2011) using Pearson correlations, here

414 ignoring uncertainty among the trend estimates.

RESULTS

415

Determining Observer-Age-Related Differences in Species Detections

416

417 *Age-related differences in OBBA detection probabilities.* — For each species, the β_2

418 estimates measure expected differences in detection abilities between the older and younger

419 observer cohorts. Negative values in this case indicate that a species is less-likely to be

420 detected by an older observer compared to a younger observer. Their average among all

421 43 species was negative (mean -0.66 ± 0.81 [SD] on the logit scale; median -0.48; Fig. 2).

422 These values were not normally-distributed, and only 4 of the 43 β_2 values were greater than

423 zero. Hence, the broad standard deviation is driven by a negative skew to the data, which are

424 almost-entirely below zero. When modeled in the GAM, the corresponding intercept term was

425 significantly negative ($P < 0.001$), which indicated a significant overall deficit in detection

426 ability among older observers, on average. On a species-specific basis, thirteen of the 43

427 species considered (30%; BAWW, BBWA, BTGW, COYE, CSWA, GCKI, NAWA, OSFL,

428 OVEN, RCKI, WIWA, YBFL, YRWA; see Appendix 1 for full names) showed 'significant'

429 declines in detectability between younger and older OBBA observers (i.e. 95% Bayesian

430 credible intervals of β_2 coefficients did not contain zero). Gender had a less-important

431 influence on detection probability; in this case, the mean effect was much closer to zero

432 (mean 0.17 ± 0.59 [SD]), and seven of the 43 (16%) species showed 'significant' effects of

433 being male on detectability. Contrary to our physiological expectations, most of the gender

434 coefficients (effect of being male) were greater than zero (34 of 43).

435 Among each of the monotone and heterogeneous species groups, and considering all peak

436 vocalization frequencies, age-related detectability differences were not significantly related to

17

437 the vocalization frequencies (GAM 'slope' smooth terms: $P = 0.297$ [monotone], $P = 0.597$

438 [heterogeneous]). However, the shape of the curve for monotone species suggested deficits

439 in detectability between 3 kHz and 6 kHz ('notch' frequencies), and beyond a threshold of

440 approximately 6 kHz ('medium' and 'high' frequencies; Fig. 2A). To test for observer-age-related

441 differences in detectability in the higher (≥ 6 kHz) frequency range exclusively, we built

442 a *post hoc*, similarly-weighted linear model predicting observer-age-related differences in

443 detectability as a function of peak vocalization frequency, using only those species with

444 monotone vocalizations above 6 kHz. This model showed a significant linear decline ($P =$

445 0.034, $n = 9$ species).

446   The sensitivity analysis tested how including observers aged 40–50 in the 'younger' age

447 group affected the relative difference in detection ability between 'younger' observers and

448 those over 50 ('older' observers). When observers between ages 40 and 50 were included

449 in the models as 'younger' participants, the the differences in detection ability between the

450 'younger' and the 'older' groups (i.e. the β_2 values) tended to diminish in magnitude (Fig.

451 3). This points to a robust effect of observer age on the detection probability of bird species.

452   *Expected counts derived from BBS data.* — Model-estimated BBS counts declined

453 significantly (GAMM smooth term $P < 0.05$) over 39 years of increasing observer age for all

454 vocalization frequency groups except Low Monotone ($P = 0.111$) and High Heterogeneous

455 species ($P = 0.085$; Fig. 4). Among the significant declines estimated, the greatest absolute

456 changes in expected counts were among the low-frequency, heterogeneous species (BCFL,

457 GCFL, Appendix 1), which decreased by 66.5%, and the medium-frequency, heterogeneous

458 species (EAKI, LEFL, NAWA, NOPA, OVEN; Appendix 1), which decreased by 59.2% over

459 the 39 years sampled. The smallest significant changes in counts were declines of 34.1%

460 among high-frequency monotone birds (BBWA, CMWA, GCKI, Appendix 1), and 34.3%

461 among notch-frequency monotone birds ($n = 18$ species; Appendix 1) over that same age

462 range (Fig. 4). The increasing uncertainty at the upper range of observer ages (Fig. 4) reflects

463 the smaller sample sizes in this area.

Species Detection Probabilities and Long-term Population Trend Estimates

Golden-crowned Kinglet case study. — Model-estimated BBS counts of the Golden-crowned Kinglet were expected to decline by nearly 7 birds per observer-route time series after 30 years of observer aging, all else being equal (Fig. 5D). Without a correction for such an effect, population trends as inferred visually from the smooth functions for calendar date appeared stable in the Northern Spruce-Hardwoods stratum (Fig. 5E), and declining in the South Pacific Rainforests and Sierra Nevada strata (Fig. 5G and I). After correcting for observer age (essentially a vector subtraction of Fig. 5D from each population smoother), inferred population trajectories became more positive. The 'corrected' Northern Spruce-Hardwoods stratum now showed a significant population increase, the South Pacific Rainforests stratum now appeared stable, and the apparent Sierra Nevada decline was less steep.

The influence of aging observers on population trend estimates. — There were significant relationships between monotone vocalization frequencies and long-term, Canada-wide population trends for each of the USGS (GAM $P = 0.048$; Fig. 6A), and CWS datasets ($P = 0.008$; Fig. 6C), where population trends were more negative among species with increasing 'medium' and 'high' (≥ 6 kHz) peak vocalization frequencies, and at the midpoint of the 'notched' range (3 to 6 kHz). By contrast, there were no significant relationships between heterogeneous vocalization frequencies and population trends (USGS $P = 0.928$; CWS $P = 0.568$; Figs. 6B and D).

As a whole, the monotone and heterogeneous patterns were also visually similar to those we observed between detection probabilities and vocalization frequencies (i.e. Fig. 2). Ignoring uncertainty in the estimated differences in detection probability (β_2), the relationship between these values (OBBA data) and estimated population trends (BBS data) were significant for both the USGS data (Pearson correlations: $r = 0.79$, $P = 0.012$; $n = 9$; Fig. 7A) and the CWS data ($r = 0.89$, $P = 0.001$; $n = 9$; Fig. 7B).

DISCUSSION

489 Using data from both detection-nondetection (OBBA) and point count (BBS) surveys; we

490 found several lines of evidence for age-related declines in bird detection abilities among

491 volunteer observers. On average, OBBA observers over age 50 had lower detection probabilities

492 compared to observers under age 40 (Fig. 2), there were more pronounced detection probability

493 differences between more distant OBBA age groups (i.e. Fig. 3), and there were near-universal

494 declines in expected BBS counts with increasing observer age (Fig. 4). Among monotone

495 species with peak vocalization frequencies exceeding 6 kHz, age-related detectability deficits

496 in the OBBA showed a significant linear decline as peak frequency increased, and detection

497 deficits may also exist at 'notched' frequencies of 3 to 6 kHz (Fig. 2). Collectively, these

498 data suggest that observer aging is an important source of error in long-term time series

499 from volunteer birdwatchers, and that common patterns of hearing loss play a role in this

500 overall process. However, the concurrent declines in detection ability we observed at other

501 frequencies (e.g. Fig. 4) suggest that other mechanisms are also involved.

502 Using real data, we found that failing to account for differences in observer detection

503 ability can underestimate population increases, and perhaps more importantly, overestimate

504 population declines (Fig. 5). We lastly found indirect evidence of such a bias in previously-published

505 population trend estimates, which tended to be lower as monotone vocalization frequencies

506 above 6 kHz increased, and towards the midpoint of the 'notched' frequency range (3 to 6

507 kHz; Fig. 6). Furthermore, we also found a tendency for species with more negative estimated

508 age-related detectability declines (estimated using an independent [OBBA] dataset) to have

509 more negative estimated BBS population trends (Fig. 7). This suggests that an uncontrolled

510 confounding effect of hearing loss and other age-related phenomena is already present in

511 population trend data.

512 The estimated declines in BBS counts we observed were consistent with data from

513 similar, previous research (Link and Sauer 1998). Here, we showed declines ranging from 34%

514 to 67% of the original counts of more than 60 species considered collectively over 39 years

20

516 (Fig. 4), whereas Link and Sauer (1998) estimated a 43% decline in Blue-Gray Gnatcatcher

517 counts among observers after 20 years (our dataset did not include the Gnatcatcher). Using

518 our approach to classifying vocalizations by peak frequency, this gnatcatcher would fall into

519 the Notch Heterogeneous category, for which we estimated a 14.1% decline in counts over

520 20 years. The smaller value predicted here may result from the large number of species

521 ($n = 25$) incorporated into this latter average. In our opinion, either value is large enough

522 to be concerning.

523   Because the declines in species detections with observer age in both the OBBA and

524 the BBS datasets occurred across most frequencies and in both heterogeneity groups –

525 for instance, the greatest modeled declines in BBS counts were for the Low and Medium

526 Heterogeneous species – we believe that multiple senescence effects, including hearing losses,

527 are at work, with hearing-loss effects being most important for species with monotone

528 vocalizations. Normal aging can involve impairments in memory, cognitive speed and vision

529 (Morris and McManus 1991). Alongside hearing impairments, these factors might each

530 contribute to greater numbers of missed detections independent of bird vocalization frequencies,

531 for instance by limiting one's abilities to (i) simultaneously detect and transcribe species calls,

532 (ii) recognize multiple, overlapping species calls, and (iii) identify non-vocal, cryptic species

533 by eye. Hence, while higher-frequency and notched monotone species might be most prone

534 to the effects of age-related hearing loss, all species are probably vulnerable to some form of

535 age-related detection decline.

536   Similarly, because behaviour and visual cues also play a role in bird detection and can

537 vary from species to species (e.g. variable species 'conspicuousness'; Stewart 1954), some

538 species with less-audible vocalizations might be detected more-often than less-conspicuous

539 species, for instance cryptic species who sing less frequently (Alldredge et al. 2007a). This

540 could explain some of the unusual patterns in the detection curves generated here (Fig. 4),

541 for instance the smaller-than-expected declines in the expected counts of High Monotone and

542 Notch Monotone species (i.e. Figs. 4C and 4G). Future controlled experiments using observers

543 of known hearing thresholds and ages, with exposure to a variety of bird vocalizations

544 of known audiological characteristics, would help to elucidate the relative importance of

545 hearing and non-hearing aging effects, as well as the interaction between these processes and

546 species-specific behaviours in the field.

547   GAM- and GAMM-based methods are relatively new to ecology (e.g. Fewster et al.

548 2000; Clarke et al. 2003; Flemming et al. 2010), but we have shown their usefulness for

549 making corrections for continuous, nonlinear covariates such as the changes in hearing ability

550 with observer age. To further minimize the influence of hearing loss on the quality of bird

551 surveys, Emlen and DeJong (1992) suggest that administrators test hearing abilities ahead of

552 time and recommend the use of hearing aids where appropriate. Especially because hearing

553 aids might not be equivalent to normal hearing, we argue that administrators should also

554 collect observer ages and information about hearing ability (once hearing aids are in place, if

555 relevant) in order to make model-based corrections, for instance using the GAMM approach

556 described here. Universal double-observer methods with younger partners (e.g. Nichols et al.

557 2000; Alldredge et al. 2006, but see Fitzpatrick *et al.* 2009), although perhaps impractical

558 in terms of the number of available skilled participants, might also help to account for this

559 source of error. A long-term solution, once the appropriate technologies become sufficiently

560 effective, portable and inexpensive, might involve a move to field recordings and interpretation

561 by a central office, rather than relying upon on-site classifications (e.g. Campbell and Francis

562 2011). Under this scheme, managers might preserve the appeal of the BBS to volunteer

563 observers/recorders if the volunteers are able to 'score' their own set of detection records

564 against those derived from such recordings. However any such protocol changes should aim

565 to be consistent throughout the survey as a whole, they must be cost-effective, and ideally,

566 they should not compromise the long-term integrity of the overall time series (e.g. Freeman

567 et al. 2007).

568   In both analyses, we controlled for many suspected observer-effects confounders by

569 excluding data. For instance, we excluded observers aged 40–50 in the analysis of OBBA

22

570 data to ensure reliable separation between 'young' and 'old' groups, improving the relative

571 accuracy of our age classfications. We also required a minimum of 10 years of service on the

572 BBS for an observer's data to be included, here to increase the likelihood that senescence

573 effects occur for all observers, so that we might measure them. While these conservative

574 approaches were appropriate for precisely determining the nature of observer senescence

575 effects, they limit the quality of the real (simultaneous) population trends that can be inferred

576 (Link and Sauer 1997). Future analyses should also explore the sensitivity of the observer-

577 and population-specific patterns we observed here to increasingly relaxed data subsetting

578 rules. In general, having surveys collect observer age data in the future would obviate any

579 future need to exclude cohorts of uncertain age, and allow models to make more-precise

580 corrections than those we have built here.

581 At a minimum, asking older or noise-exposed observers who are at risk for detection

582 errors to consider the possibility of any age-related impairments is an important step forward:

583 as with any gradual physiological change, observers over age 50 may not recognize a growing,

584 but significant personal impairment (A. G. Horn, pers. comm.), and awareness of this fact

585 alone may lead to an increased degree of self-selection in terms of opting out of surveys. For

586 instance, 75% of a sample of 253 Audubon Christmas Bird Count observers have indicated

587 a desire to remove themselves from survey duties if such an impairment was recognized

588 (Downes 2004). On the other hand, older birdwatchers are likely to be more experienced

589 and consequently more adept at detecting a wide range of rare and common species. For

590 effectively sampling entire species communities, this experience advantage may outweigh

591 early physiological deficiencies in the detection of certain species (Ramsey and Scott 1981),

592 especially when there are multiple observers per sampling unit (but see Fitzpatrick et al.

593 2009).

594 In general, our study adds to the growing body of literature demonstrating systematic,

595 long-term changes in BBS survey conditions (e.g. Betts et al. 2007; Griffith et al. 2010) that

596 must be controlled for when estimating measures of population change. We have shown

that observer age can be a significant source of error, and have illustrated some ways that survey designs and models might control for its effects. While this research focused on bird observations in particular, its implications are generalizable to other auditory wildlife surveys, for instance, anuran call counts (e.g. McClintock et al. 2010a). We hope that this research leads to improvements in long-term population trajectory inferences, while preserving the invaluable contributions made by volunteers to worldwide ecological monitoring.

ACKNOWLEDGMENTS

LITERATURE CITED

Alldredge, M. W., K. H. Pollock, and T. R. Simons. 2006. Estimating detection probabilities from multiple-observer point counts. Auk 123:1172–1182.

Alldredge, M. W., T. R. Simons, and K. H. Pollock. 2007a. Factors affecting aural detections of songbirds. Ecological Applications 17:948–955.

Alldredge, M. W., T. R. Simons, and K. H. Pollock. 2007b. A field evaluation of distance

619 measurement error in auditory avian point count surveys. Journal of Wildlife Management

620 71:2759–2766, doi:10.2193/2006-161.

621 Betts, M. G., D. Mitchell, A. W. Diamond, and J. Bety. 2007. Uneven rates of landscape

622 change as a source of bias in roadside wildlife surveys. Journal of Wildlife Management

623 71:2266–2273.

624 Bird Studies Canada, Environment Canada's Canadian Wildlife Service, Ontario Nature,

625 Ontario Field Ornithologists, and Ontario Ministry of Natural Resources. 2008. Atlas of

626 the Breeding Birds of Ontario Database [online]. URL `http://www.naturecounts.ca/`.

627 Blaustein, A. R., D. B. Wake, and W. P. Sousa. 1994. Amphibian declines - Judging stability,

628 persistence, and susceptibility of populations to local and global extinctions. Conservation

629 Biology 8:60–71.

630 Brand, A. R. 1938. Vibration frequencies of passerine bird song. Auk 55:263–268.

631 Campbell, M., and C. Francis. 2011. Using stereo microphones to evaluate observer variation

632 in North American Breeding Bird Survey (BBS) point counts. Auk 128:303–312, doi:10.

633 1525/auk.2011.10005.

634 Carver, E. 2009. Birding in the United States: A demographic and economic analysis.

635 Addendum to the 2006 Survey of Fishing, Hunting, and Wildlife-Associated Recreation.

636 Technical Report 2006-4, Arlington, VA.

637 Clarke, E. D., L. B. Spear, M. L. Mccracken, F. F. C. Marques, D. L. Borchers, S. T.

638 Buckland, and D. G. Ainley. 2003. Validating the use of generalized additive models and

639 at-sea surveys to estimate size and temporal trends of seabird populations. Journal of

640 Applied Ecology 40:278–292.

641 Cooper, C. B., W. M. Hochachka, and A. A. Dhondt. 2007. Contrasting natural experiments

642 confirm competition between house finches and house sparrows. Ecology 88:864–870.

643 Cruickshanks, K. J., D. M. Nondahl, T. S. Tweed, T. L. Wiley, B. E. K. Klein, R. Klein,

644 R. Chappell, D. S. Dalton, and S. D. Nash. 2010. Education, occupation, noise exposure

645 history and the 10-yr cumulative incidence of hearing impairment in older adults. Hearing

646 Research 264:3–9, doi:10.1016/j.heares.2009.10.008.

647 Cyr, A. 1981. Limitation and variability in hearing ability in censusing birds. Studies in

648 Avian Biology 6:327–333.

649 Downes, C. M. 2004. Results of the 2004 questionnaire for Canadian participants in the

650 Breeding Bird Survey. URL http://www.cws-scf.ec.gc.ca/nwrc-cnrf/default.asp?

651 lang=En&n=929AA800-1.

652 Dunn, E. H. 2002. Using decline in bird populations to identify needs for conservation action.

653 Conservation Biology 16:1632–1637.

654 Emlen, J. T., and M. J. DeJong. 1992. Counting birds - the problem of variable hearing

655 abilities. Journal of Field Ornithology 63:26–31.

656 Faaborg, J., R. T. Holmes, A. D. Anders, K. L. Bildstein, K. M. Dugger, S. A. Gauthreaux,

657 P. Heglund, K. A. Hobson, A. E. Jahn, D. H. Johnson, S. C. Latta, D. J. Levey, P. P.

658 Marra, C. L. Merkord, E. Nol, S. I. Rothstein, T. W. Sherry, T. S. Sillett, F. R. Thompson,

659 and N. Warnock. 2010. Conserving migratory land birds in the New World: Do we know

660 enough? Ecology 20:398–418, doi:10.1890/09-0397.1.

661 Faanes, C. A., and D. Bystrak. 1981. The role of observer bias in the North American

662 Breeding Bird Survey. Studies in Avian Biology 6:353–359.

663 Fewster, R. M., S. T. Buckland, G. M. Siriwardena, S. R. Baillie, and J. D. Wilson. 2000.

664 Analysis of population trends for farmland birds using generalized additive models. Ecology

665 81:1970–1984.

666 Fitzpatrick, M. C., E. L. Preisser, A. M. Ellison, and J. S. Elkinton. 2009. Observer bias and

667   the detection of low-density populations. Ecological Applications 19:1673–1679, doi:10.

668   1890/09-0265.1.

669 Flemming, J. M., E. Cantoni, C. Field, and I. McLaren. 2010. Extracting long-term patterns

670   of population changes from sporadic counts of migrant birds. Environmetrics 21:482–492,

671   doi:10.1002/env.998.

672 Freeman, S. N., D. G. Noble, S. E. Newson, and S. R. Baillie. 2007. Modelling population

673   changes using data from different surveys: the common birds census and the breeding bird

674   survey. Bird Study 54:61–72.

675 Gates, G. A., and J. H. Mills. 2005. Presbycusis. Lancet 366:1111–1120.

676 Gates, G. A., P. Schmid, S. G. Kujawa, B. H. Nam, and R. D'Agostino. 2000. Longitudinal

677   threshold changes in older men with audiometric notches. Hearing Research 141:220–228.

678 Gelman, A. 2006. Prior distributions for variance parameters in hierarchical models

679   (Comment on an Article by Browne and Draper). Bayesian Analysis 1:515–533.

680 Gelman, A. 2008. Scaling regression inputs by dividing by two standard deviations. Statistics

681   in Medicine 27:2865–2873, doi:10.1002/sim.3107.

682 Gelman, A., A. Jakulin, M. G. Pittau, and Y. S. Su. 2008. A weakly informative default

683   prior distribution for logistic and other regression models. Annals of Applied Statistics

684   2:1360–1383, doi:10.1214/08-AOAS191.

685 Gelman, A., Y. Su, M. Yajima, J. Hill, M. G. Pittau, J. Kerman, and T. Zheng. 2010. arm:

686   Data analysis using regression and multilevel/hierarchical models. URL `http://CRAN.`

687   `R-project.org/package=arm`.

688 Goffredo, S., F. Pensa, P. Neri, A. Orlandi, M. Scola Gagliardi, A. Velardi, C. Piccinetti, and

689 F. Zaccanti. 2010. Unite research with what citizens do for fun: "Recreational monitoring"

690 of marine biodiversity. Ecological Applications 20:2170–2187, doi:10.1890/09-1546.

691 Greenberg, R., and S. Droege. 1999. On the decline of the rusty blackbird and the use of

692 ornithological literature to document long-term population trends. Conservation Biology

693 13:553–559.

694 Griffith, E. H., J. R. Sauer, and J. A. Royle. 2010. Traffic effects on bird counts on North

695 American Breeding Bird Survey routes. Auk 127:387–393, doi:10.1525/auk.2009.09056.

696 International Organization for Standardization. 2000. Acoustics- Statistical distribution of

697 hearing thresholds as a function of age (ISO 7029:2000). Geneva, Switzerland.

698 Johnson, D. H. 2008. In defense of indices: The case of bird surveys. Journal of Wildlife

699 Management 72:857–868.

700 Kendall, W. L., B. G. Peterjohn, and J. R. Sauer. 1996. First-time observer effects in the

701 North American Breeding Bird Survey. Auk 113:823–829.

702 Kéry, M., J. A. Royle, H. Schmid, M. Schaub, B. Volet, G. Hafliger, and N. Zbinden. 2010.

703 Site-occupancy distribution modeling to correct population-trend estimates derived from

704 opportunistic observations. Conservation Biology 24:1388–1397, doi:10.1111/j.1523-1739.

705 2010.01479.x.

706 Klimkiewicz, M. K., and C. S. Robbins. 1978. Standard abbreviations for common names of

707 birds. North American Bird Bander 3:16–25.

708 Kremen, C., K. S. Ullman, and R. W. Thorp. 2011. Evaluating the quality of

709 citizen-scientist data on pollinator communities. Conservation Biology 25:607–617, doi:10.

710 1111/j.1523-1739.2011.01657.x.

711 La Rouche, G. P. 2001. Birding in the United States: A demographic and economic analysis.

712   Addendum to the 2001 Survey of Fishing, Hunting, and Wildlife-Associated Recreation.

713   Technical Report 2001-1, Washington, DC.

714   La Sorte, F. A., and M. L. McKinney. 2007. Compositional changes over space and time

715   along an occurrence-abundance continuum: Anthropogenic homogenization of the North

716   American avifauna. Journal of Biogeography 34:2159–2167.

717   LaDeau, S. L., A. M. Kilpatrick, and P. P. Marra. 2007. West Nile virus emergence and

718   large-scale declines of North American bird populations. Nature 447:710–U13.

719   Link, W. A., and J. R. Sauer. 1997. New approaches to the analysis of population trends in

720   land birds: Comment. Ecology 78:2632–2634.

721   Link, W. A., and J. R. Sauer. 1998. Estimating population change from count data:

722   Application to the North American Breeding Bird Survey. Ecological Applications

723   8:258–268.

724   Link, W. A., and J. R. Sauer. 2002. A hierarchical analysis of population change with

725   application to Cerulean Warblers. Ecology 83:2832–2840.

726   Link, W. A., and J. R. Sauer. 2007. Seasonal components of avian population change: joint

727   analysis of two large-scale monitoring programs. Ecology 88:49–55.

728   Link, W. A., J. R. Sauer, and D. K. Niven. 2008. Combining breeding bird survey and

729   Christmas Bird Count data to evaluate seasonal components of population change in

730   northern bobwhite. Journal of Wildlife Management 72:44–51, doi:10.2193/2007-299.

731   Lotz, A., and C. R. Allen. 2007. Observer bias in anuran call surveys. Journal of Wildlife

732   Management 71:675–679, doi:10.2193/2005-759.

733   Lunn, D., D. Spiegelhalter, A. Thomas, and N. Best. 2009. The BUGS project: Evolution,

734   critique and future directions. Statistics in Medicine 28:3049–3067, doi:10.1002/sim.3680.

29

735 MacKenzie, D. I., J. D. Nichols, N. Sutton, K. Kawanishi, and L. L. Bailey. 2005. Improving

736 inferences in population studies of rare species that are detected imperfectly. Ecology

737 86:1101–1113.

738 Mayfield, H. 1966. Hearing loss and bird song. Living Bird 5:167–175.

739 McClintock, B. T., L. L. Bailey, K. H. Pollock, and T. R. Simons. 2010a. Experimental

740 investigation of observation error in anuran call surveys. Journal of Wildlife Management

741 74:1882–1893, doi:10.2193%2F2009-321.

742 McClintock, B. T., L. L. Bailey, K. H. Pollock, and T. R. Simons. 2010b. Unmodeled

743 observation error induces bias when inferring patterns and dynamics of species occurrence

744 via aural detections. Ecology 91:2446–2454, doi:10.1890/09-1287.1.

745 McLaren, A. A., and M. D. Cadman. 1999. Can novice volunteers provide credible data for

746 bird surveys requiring song identification? Journal of Field Ornithology 70:481–490.

747 Miller, D. A., J. D. Nichols, B. T. McClintock, E. H. C. Grant, L. L. Bailey, and L. A.

748 Weir. 2011. Improving occupancy estimation when two types of observational error occur:

749 Non-detection and species misidentification. Ecology 92:1422–1428, doi:10.1890/10-1396.1.

750 Morris, J. C., and D. Q. McManus. 1991. The neurology of aging - Normal versus pathological

751 change. Geriatrics 46:47–54.

752 Nichols, J. D., J. E. Hines, J. R. Sauer, F. W. Fallon, J. E. Fallon, and P. J. Heglund. 2000. A

753 double-observer approach for estimating detection probability and abundance from point

754 counts. Auk 117:393–408.

755 Nichols, J. D., L. Thomas, and P. B. Conn. 2009. Inferences about landbird abundance from

756 count data: Recent advances and future directions. In D. L. Thomson, E. G. Cooch, and

757 M. J. Conroy, editors, Environmental and Ecological Statistics, volume 3, Springer US,

758 pages 201–235, doi:10.1007/978-0-387-78151-8_9.

759 Nondahl, D. A., X. Y. Shi, K. J. Cruickshanks, D. S. Dalton, T. S. Tweed, T. L. Wiley, and

760   L. L. Carmichael. 2009. Notched audiograms and noise exposure history in older adults.

761   Ear and Hearing 30:696–703.

762 Osei-Lah, V., and L. H. Yeoh. 2010. High frequency audiometric notch: An outpatient clinic

763   survey. International Journal of Audiology 49:95–98, doi:10.3109/14992020903300423.

764 Pearson, J. D., C. H. Morrell, S. Gordonsalant, L. J. Brant, E. J. Metter, L. L. Klein, and

765   J. L. Fozard. 1995. Gender differences in a longitudinal-study of age-associated hearing

766   loss. Journal of the Acoustical Society of America 97:1196–1205.

767 Peterjohn, B. G. 1994. The North American Breeding Bird Survey. Birding 26:386–398.

768 Peters, D. P. 2010. Accessible ecology: synthesis of the long, deep, and broad. Trends in

769   Ecology & Evolution 25:592–601.

770 R Development Core Team. 2011. R: A Language and Environment for Statistical Computing.

771   Vienna, Austria, URL `http://www.R-project.org`.

772 Ramsey, F. L., and J. M. Scott. 1981. Tests of hearing ability. Studies in Avian Biology

773   6:341–345.

774 Royle, J. A., and R. M. Dorazio. 2009. Hierarchical modeling and inference in ecology: The

775   analysis of data from populations, metapopulations and communities. Academic Press

776   (Elsevier).

777 Royle, J. A., and M. Kéry. 2007. A Bayesian state-space formulation of dynamic occupancy

778   models. Ecology 88:1813–1823.

779 Royle, J. A., and W. A. Link. 2006. Generalized site occupancy models allowing for false

780   positive and false negative errors. Ecology 87:835–841.

781 Sauer, J. R., J. E. Fallon, and R. Johnson. 2003. Use of North American Breeding Bird

782 Survey data to estimate population change for bird conservation regions. Journal of Wildlife

783 Management 67:372–389.

784 Sauer, J. R., and W. A. Link. 2011. Analysis of the North American Breeding Bird Survey

785 using hierarchical models. Auk 128:87–98, doi:10.1525/auk.2010.09220.

786 Sauer, J. R., G. W. Pendleton, and B. G. Peterjohn. 1996. Evaluating causes of population

787 change in North American insectivorous songbirds. Conservation Biology 10:465–478.

788 Sauer, J. R., B. G. Peterjohn, and W. A. Link. 1994. Observer differences in the North

789 American Breeding Bird Survey. Auk 111:50–62.

790 Silvertown, J. 2009. A new dawn for citizen science. Trends in Ecology & Evolution

791 24:467–471, doi:10.1016/j.tree.2009.03.017.

792 Simons, T. R., M. W. Alldredge, K. H. Pollock, and J. M. Wettroth. 2007. Experimental

793 analysis of the auditory detection process on avian point counts. Auk 124:986–999.

794 Stewart, P. A. 1954. The value of Christmas Bird Counts. Wilson Bulletin 66:184–195.

795 Thomas, L., and K. Martin. 1996. The importance of analysis method for breeding bird

796 survey population trend estimates. Conservation Biology 10:479–490.

797 Toppila, E., I. Pyykk, and J. Starck. 2001. Age and noise-induced hearing loss. Scandinavian

798 Audiology 30:236–244, doi:10.1080/01050390152704751.

799 Wiedner, D., and P. Kerlinger. 1990. Economics of birding: A national survey of active

800 birders. American Birds 44:209–213.

801 Wiley, T. L., R. Chappell, L. Carmichael, D. A. Nondahl, and K. J. Cruickshanks. 2008.

802 Changes in hearing thresholds over 10 years in older adults. Journal of the American

803 Academy of Audiology 19:281–292, doi:10.3766/jaaa.19.4.2.

804 Wilson, S., S. LaDeau, A. Tottrup, and P. Marra. 2011. Range-wide effects of breeding and

805   non-breeding season climate on the abundance of a Neotropical migrant songbird. Ecology

806   92:1789–1798, doi:10.1890/10-1757.1.

807 Wood, S. N. 2006. Generalized Additive Models: An Introduction with R. Chapman and

808   Hall, New York.

809 Wood, S. N. 2011. gamm4: Generalized additive mixed models using mgcv and lme4. URL

810   http://CRAN.R-project.org/package=gamm4.

811 # FIGURE LEGENDS

Figure 1. International standard for normal changes in hearing thresholds (median values) at standard pure-tone test frequencies (1, 4, 6, and 8 kHz) among (A) men and (B) women of increasing age. Shaded areas are 95% quantiles. Curves are derived from models specified in International Organization for Standardization (2000).

Figure 2. Logit-scale difference in species detection probability between an observer over age 50 and an observer under age 40 ('β_2' coefficients), determined from species-specific hierarchical occupancy models of data from the Atlas of the Breeding Birds of Ontario, as a function of each species' peak vocalization frequency, and grouped by vocalization variability ('Monotone' [A] and 'Heterogeneous' [B]). Smoothed curves are GAM fits, weighted by the inverse of the variance of each datapoint (uncertainty displayed as 95% credible interval lines here), plus the model intercept. Shaded areas are 95% pointwise confidence bands about the smooth term and the model intercept. Dotted reference lines are plotted at $y = 0$ (no difference in detection between younger and older observers) and at 6 kHz (the threshold for "medium" frequencies, as defined in the text).

Figure 3. Sensitivity of observer-age-related differences in species detection probabilities ($`\beta_2`$') to the age structure of the old and young age groups in the modeled data. Coefficients were generated from identical occupancy models using data that either lacked the 40–50 age group as part of its 'young' category (principal modeling approach; x-axis) or included these observers (y-axis). The dotted reference line has a slope of 1. When the additional group of middle-aged observers are included in the 'under-50' category (y-axis), the observer-age-related differences in species detectability are pushed closer to zero.

Figure 4. Estimated proportional changes in BBS counts with increasing minimum observer age, relative to the count during the first year of BBS service, grouped by species vocalization groups. Vocalization groups reflect the peak frequency of a typical set of vocalizations for that species, and whether these vocalizations tend to feature a single sound ("Monotone") or a highly-variable set of sounds ("Heterogeneous"). Standard abbreviations for species names (Klimkiewicz and Robbins 1978) belonging to each vocalization group are listed on each panel (also see Appendix 1). Shaded areas are 95% pointwise confidence bands.

Figure 5. Case study of Golden-crowned Kinglet (*Regulus satrapa*) BBS counts, modeled as GAMMs with and without corrections for observer age. Panels A through C show raw data (with loess curves) for a random subset of observer-within-route time series within each of 3 selected physiographic strata. Panels E, G and I show modeled 'population' (calendar year) trends, correcting for first-year and among-observer/route effects. Panels F, H and J show 'population' trends after making a correction for observer age (shown explicitly in Panel D) in place of the first-year correction. Shaded areas are 95% pointwise confidence bands.

Figure 6. Additive model of Canada-wide population trends calculated by the United States Geological Service (1966 to 2009; panels A and B), and by the Canadian Wildlife Service (1970 to 2009; panels C and D), as a function of each species' peak vocalization frequency (pitch), modeled separately for species with largely single-frequency vocalizations ('Monotone'; panels A and C) and for species with highly-variable vocalization frequencies ('Heterogeneous'; panels B and D). Shaded areas are 95% pointwise confidence bands about the model smooth term plus the model intercept.

Figure 7. Observer-age-related differences in species detection probabilities ('β_2' coefficients) estimated by the OBBA occupancy models plotted against (A) USGS and (B) CWS long-term population trend estimates for Canada, for those species common to the two analyses and having monotone vocalizations of medium ($\geq$ 6 kHz) or high ($\geq$ 7 kHz) frequencies. Solid lines correspond to linear regression fits which ignore uncertainty in the detection probability (x-axis) values. Both regression slopes are significantly ($P < 0.05$) different from zero. 'BAWW' is the Black-and-white Warbler (*Mniotilta varia*). This figure combines information from Figures 2 and 6.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

812

Figures

Figure 1

Figure 2

36

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
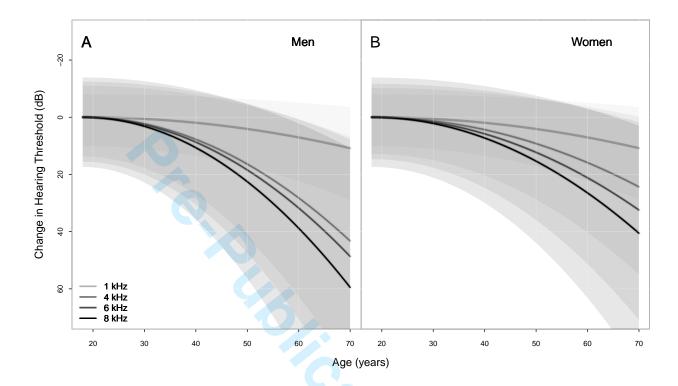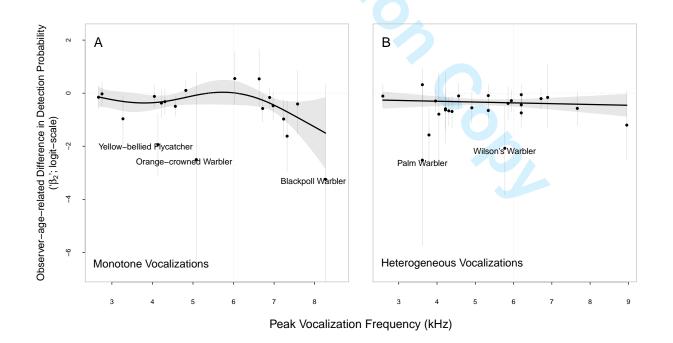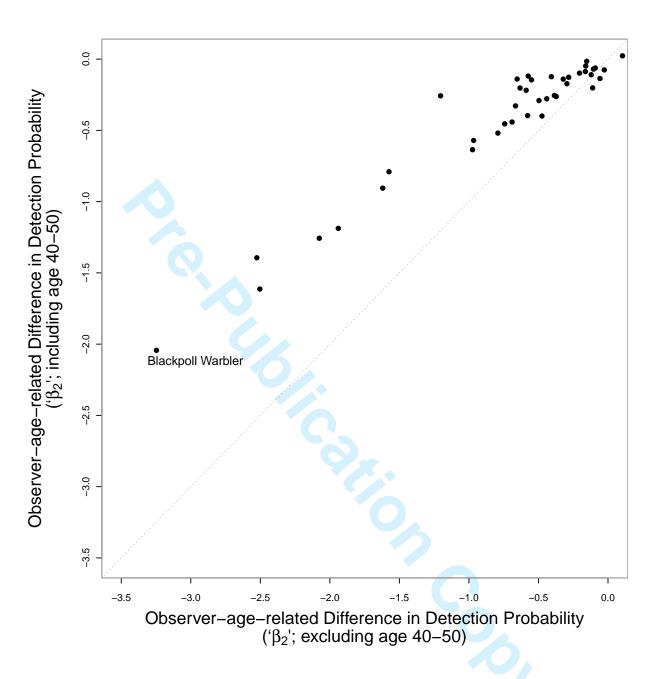36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure 3

Figure 4

Figure 5

Figure 6

Figure 7

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
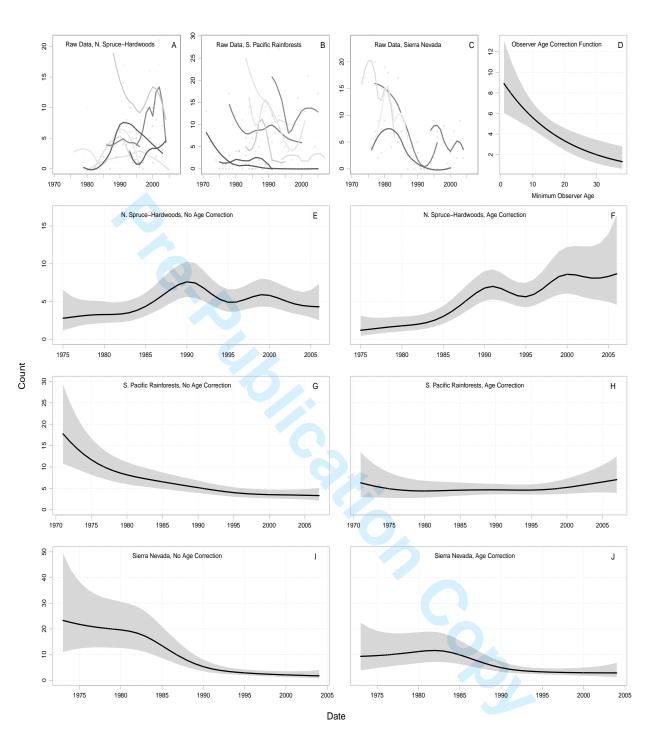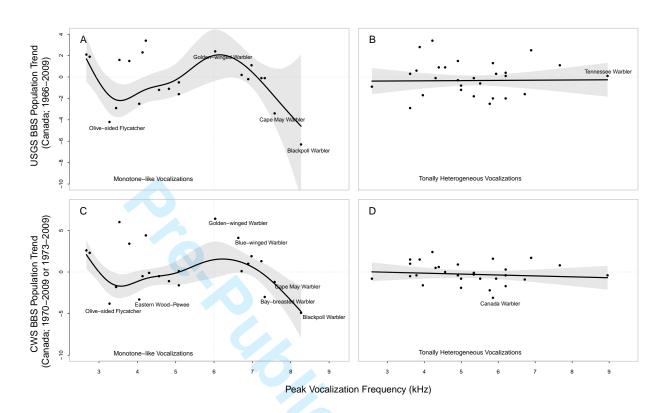59
60

813
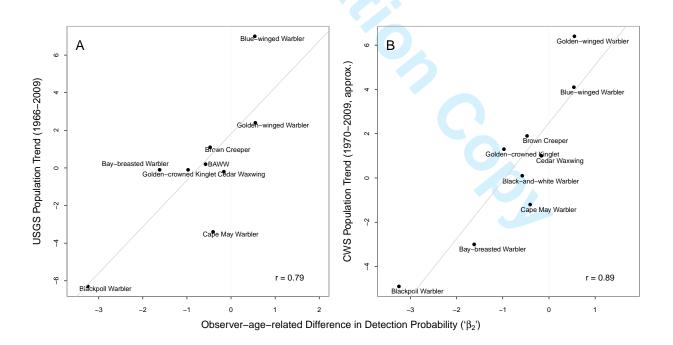
Appendices

Appendix 1. Table of species used in the various hearing-loss analyses. 'USGS' and 'CWS' refer to the respective agencies' population trend analyses. Standard species abbreviations are taken from Klimkiewicz and Robbins (1978). Data source is indicated by an asterisk in the corresponding column. Vocalization frequency information, including peak vocalization frequency (Hz) and power spectrum standard deviation ('SD', as an index of call heterogeneity) are also provided. Frequency range and heterogeneity classifications are also provided, where *low* frequencies are less than 3 kHz, *'notch'* frequencies (corresponding to the audiometric notch related to noise-induced hearing loss) are between 3 kHz and less than 6 kHz, *medium* frequencies are between 6 and less than 7 kHz, and *high* frequency calls exceed 7 kHz. Heterogeneous vocalizations are in the upper 50% quantile of standard deviation values for a group of species that includes 19 additional, unmodeled species (not shown).

Species	Abbrev.	OBBA	BBS	USGS	CWS	Peak Freq. (Hz)	SD	Class
Red-breasted Nuthatch	RBNU	*	*	*	*	2670	514.22	Low Monotone
White-breasted Nuthatch	WBNU	*	*	*	*	2756	329.45	Low Monotone
Brown-crested Flycatcher	BCFL		*			2412	717.27	Low Heterogeneous
Great Crested Flycatcher	GCFL	*	*		*	2584	821.43	Low Heterogeneous
Alder Flycatcher	ALFL	*	*		*	4307	646.83	Notch Monotone
Ash-throated Flycatcher	ATFL		*	*		3101	319.57	Notch Monotone
Black-throated Gray Warbler	BTGW				*	5082	618.78	Notch Monotone
Brown-headed Nuthatch	BHNU		*	*		4393	576.00	Notch Monotone

Continued on next page

42

			Appendix 1, continued					
Species	Abbrev.	OBBA	BBS	USGS	CWS	Peak Freq. (Hz)	SD	Class
Cassin's Kingbird	CAKI			*		3273	606.45	Notch Monotone
Common Yellowthroat	COYE	*	*	*	*	4565	593.58	Notch Monotone
Eastern Phoebe	EAPH	*	*	*	*	4823	432.01	Notch Monotone
Eastern Wood-Pewee	EAWP	*	*	*	*	4048	475.55	Notch Monotone
Grace's Warbler	GRWA			*		3876	489.64	Notch Monotone
Gray Flycatcher	GRFL			*		3790	678.03	Notch Monotone
Kentucky Warbler	KEWA		*	*		4910	684.07	Notch Monotone
Lucy's Warbler	LUWA		*	*		5512	630.58	Notch Monotone
Olive-sided Flycatcher	OSFL	*	*	*	*	3273	523.34	Notch Monotone
Orange-crowned Warbler	OCWA	*	*	*	*	5082	562.35	Notch Monotone
Pine Warbler	PIWA	*	*	*	*	4221	532.83	Notch Monotone
Pygmy Nuthatch	PYNU		*	*	*	3790	317.78	Notch Monotone
Say's Phoebe	SAPH		*	*	*	3531	453.42	Notch Monotone
Scissor-tailed Flycatcher	STFL		*			3704	623.82	Notch Monotone
Swainson's Warbler	SWWA		*	*		4996	660.03	Notch Monotone
Vermilion Flycatcher	VEFL		*	*		4048	620.00	Notch Monotone
Western Wood-Pewee	WWPE		*	*	*	3445	405.60	Notch Monotone
Yellow-bellied Flycatcher	YBFL	*	*		*	4134	646.14	Notch Monotone

Appendix 1, continued								
Species	Abbrev.	OBBA	BBS	USGS	CWS	Peak Freq. (Hz)	SD	Class
Acadian Flycatcher	ACFL		*	*		4823	755.32	Notch Heterogeneous
American Redstart	AMRE	*	*	*	*	5943	915.55	Notch Heterogeneous
Black-throated Blue Warbler	BTBW	*	*		*	4221	713.75	Notch Heterogeneous
Black-throated Green Warbler	BTGW	*			*	4393	905.87	Notch Heterogeneous
Black Phoebe	BLPH		*	*		5082	951.26	Notch Heterogeneous
Canada Warbler	CAWA	*	*	*	*	5857	830.75	Notch Heterogeneous
Cerulean Warbler	CRWA	*	*	*		4221	879.96	Notch Heterogeneous
Chestnut-sided Warbler	CSWA	*	*	*	*	5340	1101.86	Notch Heterogeneous
Connecticut Warbler	COWA			*	*	4996	1100.18	Notch Heterogeneous
Dusky Flycatcher	DUFL		*	*	*	4910	763.74	Notch Heterogeneous
Hammond's Flycatcher	HAFL		*	*	*	5857	973.81	Notch Heterogeneous
Hermit Warbler	HEWA		*	*		5168	919.93	Notch Heterogeneous
Hooded Warbler	HOWA	*	*	*		4048	761.85	Notch Heterogeneous
Louisiana Waterthrush	LOWA		*	*		4565	796.63	Notch Heterogeneous
MacGillivray's Warbler	MGWA		*	*	*	4996	708.54	Notch Heterogeneous
Magnolia Warbler	MAWA	*	*	*	*	4910	1283.08	Notch Heterogeneous
Mourning Warbler	MOWA	*	*	*	*	3962	891.97	Notch Heterogeneous
Northern Waterthrush	NOWA	*	*	*	*	4565	1111.92	Notch Heterogeneous

44

Appendix 1, continued								
Species	Abbrev.	OBBA	BBS	USGS	CWS	Peak Freq. (Hz)	SD	Class
Palm Warbler	PAWA	*	*	*	*	3618	840.98	Notch Heterogeneous
Prairie Warbler	PRWA		*	*		5340	855.07	Notch Heterogeneous
Ruby-crowned Kinglet	RCKI	*	*	*	*	3790	1144.54	Notch Heterogeneous
Townsend's Warbler	TOWA		*	*	*	5512	860.07	Notch Heterogeneous
Virginia's Warbler	VJWA		*	*		5082	696.80	Notch Heterogeneous
Western Kingbird	WEKI		*	*	*	3618	1004.12	Notch Heterogeneous
Willow Flycatcher	WIFL	*	*		*	3618	693.10	Notch Heterogeneous
Wilson's Warbler	WIWA	*		*	*	5771	1122.54	Notch Heterogeneous
Yellow-breasted Chat	YBCH		*	*	*	3876	1212.40	Notch Heterogeneous
Yellow-rumped Warbler	YRWA	*		*	*	4307	832.30	Notch Heterogeneous
Yellow-throated Warbler	YTWA		*	*		5857	732.75	Notch Heterogeneous
Yellow Warbler	YEWA	*	*	*	*	5340	889.48	Notch Heterogeneous
Black-and-white Warbler	BAWW	*	*	*	*	6718	663.48	Medium Monotone
Blue-winged Warbler	BWWA	*	*	*	*	6632	591.23	Medium Monotone
Brown Creeper	BRCR	*	*	*	*	6977	679.02	Medium Monotone
Cedar Waxwing	CEWA	*	*	*	*	6891	314.65	Medium Monotone
Golden-winged Warbler	GWWA	*	*	*	*	6029	424.40	Medium Monotone
Worm-eating Warbler	WEWA		*	*		6546	514.35	Medium Monotone

45

				Appendix 1, continued				
Species	Abbrev.	OBBA	BBS	USGS	CWS	Peak Freq. (Hz)	SD	Class
Cordilleran Flycatcher	COFL		*			6460	973.74	Medium Heterogeneous
Eastern Kingbird	EAKI	*	*	*	*	6202	1171.83	Medium Heterogeneous
Least Flycatcher	LEFL	*	*	*	*	6718	1276.71	Medium Heterogeneous
Nashville Warbler	NAWA	*	*	*	*	6202	961.56	Medium Heterogeneous
Northern Parula	NOPA	*	*	*	*	6891	786.68	Medium Heterogeneous
Ovenbird	OVEN	*	*	*	*	6202	991.24	Medium Heterogeneous
Bay-breasted Warbler	BBWA	*	*	*	*	7321	490.08	High Monotone
Blackpoll Warbler	BPWA	*		*	*	8269	257.59	High Monotone
Cape May Warbler	CMWA	*	*	*	*	7580	375.51	High Monotone
Golden-crowned Kinglet	GCKI	*	*	*	*	7235	680.00	High Monotone
Blackburnian Warbler	BLWA	*		*	*	7666	828.64	High Heterogeneous
Prothonotary Warbler	POWA		*	*		7494	1213.31	High Heterogeneous
Tennessee Warbler	TEWA	*	*	*	*	8958	1216.47	High Heterogeneous

46

Appendix 2. Supplementary Figures

Figure A1. Age distributions among bird surveyor gender cohorts. Panel A shows a beanplot (Kampstra, P. 2008. Journal of Statistical Software. 28:1–9) of the distribution of age ranges among a small sample of BBS observers, based upon demographic information collected by an unrelated internet-based survey of birdwatcher observer effects (Farmer et al. 2012. Auk 129:76–86. doi:10.1525/auk.2012.11129). Tick mark lengths correspond to observer abundance at each age range; the dotted line is the overall mean, solid lines are group means. Panel B shows a barplot of the genders and estimated ages of those OBBA observers determined for the current study.

Figure A2. Examples of audiological power spectra corresponding to monotone (Blackpoll Warbler, *Setophaga striata*) and heterogeneous (Tennessee Warbler, *Oreothlypis peregrina*) vocalizations. The modified version displayed here presents the power as a linear-scale version of decibel values for each of a continuous range of frequency bins. Monotone vocalizations tend to feature a single or narrow range of frequencies, whereas heterogeneous vocalizations feature a wide range of sounds. Peak frequency values (kHz) and SD values (as a measure of heterogeneity) are listed for each species.

Figure A3.　Probability density functions of selected prior distributions suggested or recommended by Gelman et al. (2008) for logistic Bayesian models, along with the $N(0, \sqrt{10})$ prior distribution used in the detectability difference models here (Appendix 4).

Appendix 3. Vocalization Heterogeneity

For each species, we obtained an audio recording of its typical vocalizations (calls and songs) from the Macaulay Library at the Cornell Laboratory of Ornithology (`http://macaulaylibrary.org`) and generated power spectra from each recording using the free software Audacity (Beta 1.3; `http://audacity.sourceforge.net/`). Power spectra display the total energy expended during an audio sample (dB) for each of a contiguous range of narrow frequency bins (i.e. 2.00–2.08 kHz, 2.081–2.160 kHz; Fig. A2 in Appendix 2). To reduce the effects of any background noise, we converted the log-scale power (dB) values to a linear equivalent. We then noted the peak frequency, defined as the upper bound of the frequency bin with the highest power. With this approach, the length of the recordings and the number of vocalizations featured in each recording were unimportant, as the power spectra considered the power and frequencies of all sounds present on each recording collectively.

By convention, sound intensities (power) are scored on the (logarithmic) decibel scale, which recognizes that human ears most readily distinguish changes in intensity along such an axis (Mayfield 1966). Converting a set of sound intensities to linear scales would tend to de-emphasize softer notes and highlight differences only among sounds of higher intensities. In our case, this approach was highly-appropriate for comparing vocalization variability in that it tended to downplay any background noises present on a given audio track and emphasize only the dominant singing and calling notes of a given species. Accordingly, to classify vocalizations into 'monotone' and 'heterogeneous' groups according to the variability of frequencies they contain, we first rescaled and linearized the log-scale decibel values within each power spectrum using the formula:

$$RelPower_i = 10^{(Power_i - Power_{max}) \cdot 0.1} \tag{1}$$

where $(Power_i - Power_{max})$ corresponds to the (negative) linear difference on the decibel scale between a given power value and the spectrum's maximum power value for $1, \ldots, i$ frequency

839   bins. This function converts all decibel values to a scale from 0 to 1, where 1 equals the

840   maximum power output, and it reflects linear-scale power differences (i.e. non-decibel values)

841   between any given value and the maximum value. We then treated these power spectra as

842   histograms and determined the standard deviation of these 'distributions' as a measure of

843   their acoustic variability.

844 **Appendix 4. Hierarchical Occupancy Model Structure**

845 The occupancy component of the models for each species was specified as:

$$z_i \sim Bernoulli(\psi_i) \tag{2}$$

846

$$logit(\psi_i) = A_0 + A_1 \cdot \zeta_i \tag{3}$$

847 for $i = 1, \ldots, 1212$ OBBA (atlas) squares, and where z_i corresponds to the unobserved
848 true occupancy state of a given (second-atlas) atlas square (i.e. 0 or 1), $P(z_i = 1) =$
849 ψ_i (the occupancy probability for atlas square i), and ζ_i is a dummy variable indicating
850 detection/nondetection (i.e. 0, 1) of a species by any observer in square i in the first atlas
851 (1981-1985). A_0 and A_1 are logit-scale intercept and first-year occupancy parameters. Data
852 used to determine ζ_i were derived from a set of 1,325 total observers from the first OBBA.

853 The detection component of the occupancy models for each species was specified as:

$$logit(p_{ij}) = \beta_1 \cdot \theta_{ij} + b_{obs_j} \tag{4}$$

854

$$b_{obs_j} = \beta_0 + \beta_2 \cdot Over50_j + \beta_3 \cdot Male_j + \epsilon_j \tag{5}$$

855 for $i = 1, \ldots, 1212$ atlas squares and $j = 1, \ldots, 350$ observers (or fewer, depending on the
856 species being modeled), and where p_{ij} is the detection probability at square i for observer
857 j, θ_{ij} is the natural log of effort, in party-hours, at square i by observer j, β_1 is the effort
858 effect, and b_{obs_j} describes the observer effects. Among these observer effects (equation 5),
859 β_0 is an intercept term, β_2 is the age (over-50 vs. under-40) effect, β_3 is the effect of being
860 male, and ϵ_j is mean-zero, normally-distributed error about the observer effect, with the
861 uniformly-distributed variance of this error estimated from the data (see discussion of priors,
862 below). $Over50_j$ and $Male_j$ are dummy variables (0 or 1) indicating whether an observer is
863 over age 50 (vs. under age 40), and whether that observer is male (vs. female).

864 The occupancy and detection models are combined in the overall hierarchy, which

incorporates observed detections Y_{ij}:

$$\mu_{ij} = z_i \cdot p_{ij} \tag{6}$$

$$Y_{ij} \sim Bin(N_{ij}, \mu_{ij}) \tag{7}$$

where Y_{ij}, the observed number of detections in square i for observer j is binomially distributed with probability of success μ_{ij} (the unconditional detection probability) for N_{ij} trials (i.e. the number of years during which an atlas square i was visited by observer j, which ranged from 2 to 5 detection-years).

All parameters in the hierarchical model (A_0, A_1, β_0, β_1, β_2, β_3) were assigned minimally-informative Bayesian priors suitable for logistic regression models, which in most cases need not estimate absolute values greater than 5 (Gelman et al. 2008). We specifically used normally-distributed priors of standard deviation 3.16 ($\sqrt{10}$), which, upon visual inspection of the density function, are distributed roughly similarly to the Cauchy prior of scale 2.5 recommended by Gelman et al. (2008, and see Fig. A3 in Appendix 2) for this type of model – and which WinBUGS cannot directly simulate. We also considered other priors based upon the t-distribution that are also suggested by Gelman (2008); however, a sensitivity analysis showed that these priors (t_7 with scale parameters 2.5 and 10 for the predictors and intercept terms, respectively) had slower convergence rates, a narrower range of absolute parameter estimates, and lower effective sample sizes on average. This was a good indication that the normal priors were superior for our purposes.

The ϵ_j were assigned a normal, mean-zero prior with variance estimated from the data with a mean-zero, uniform prior (Gelman 2006) of standard deviation 10, which was again consistent with the range of parameter values expected in most logistic regressions (Gelman et al. 2008). We used enough iterations in WinBUGS to achieve convergence of 3 Markov chains (with a burn-in of one half of the total), requiring that Gelman-Rubin Rhat statistics for all parameters be less than or equal to 1.1 to infer convergence. We also verified the

889 performance of this model structure using simulated datasets (see electronic supplement for

890 example code).

891 **Appendix 5. Detailed Methods for Modeling Changes in BBS Count Data with**

892 **Increasing Observer Age**

893 To keep the more heavily-sampled species, observers or strata from having a disproportionate

894 influence in our aggregated analysis, we modeled our BBS dataset over multiple stages using

895 GAMMs. First, we modeled mean BBS counts for each species separately as overdispersed

896 Poisson functions of both observer age and calendar year, correcting for differences among

897 observers and survey routes as mean-zero, normally-distributed random intercepts. We

898 used a cubic regression spline smooth term, chosen over thin-plate regression splines for

899 computational efficiency reasons (Wood 2006), for each of the observer age and calendar year

900 (i.e. population) effects, where the calendar year effects were smoothed separately for each

901 stratum. The structure for each species-specific model was as follows:

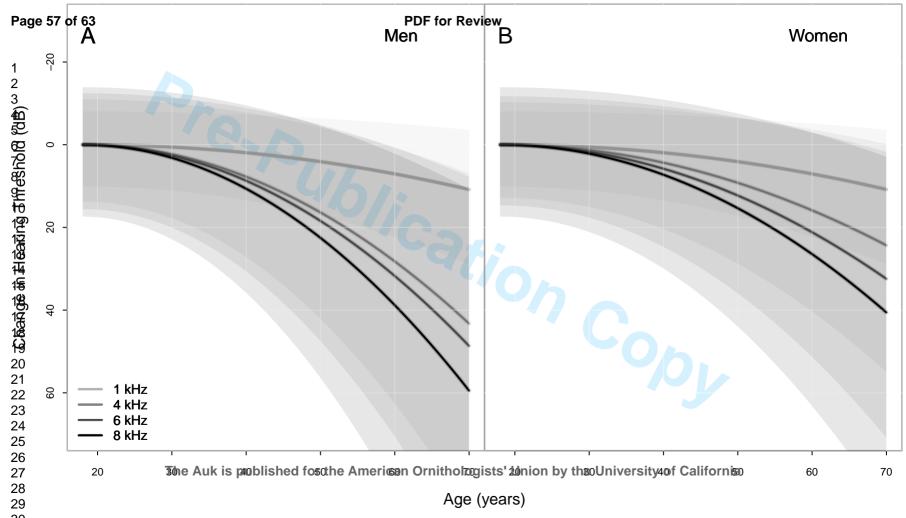$$\log(y_{i(j)kl}) = f_1(\tau_{kl}) + f_2(l)_j + \theta_k + \lambda_{i(j)k} + \sigma_{i(j)kl} \tag{8}$$

902 for $i = 1, \ldots, I$ routes within stratum $j = 1, \ldots, J$, $k = 1, \ldots, K$ observers, and $l =$

903 $1, \ldots, L$ calendar years since 1969, and where $y_{i(j)kl}$ is the number of birds detected on a

904 route i in stratum j by observer k during year l, $f_1()$ and $f_2()_j$ are cubic spline smooth

905 functions estimating age effects across the whole survey and population-related effects for

906 physiographic stratum j, respectively, τ_{kl} is the (minimum) age of observer k in year l, θ_k are

907 mean-zero, normally-distributed random intercepts for each observer, $\lambda_{i(j)k}$ are mean-zero,

908 normally-distributed random intercepts for each observer at a route-within-stratum, $\sigma_{i(j)kl}$

909 is mean-zero, normally-distributed overdispersion error, and where datapoints collected by a

910 given observer were weighted according to the inverse of the number of routes conducted by

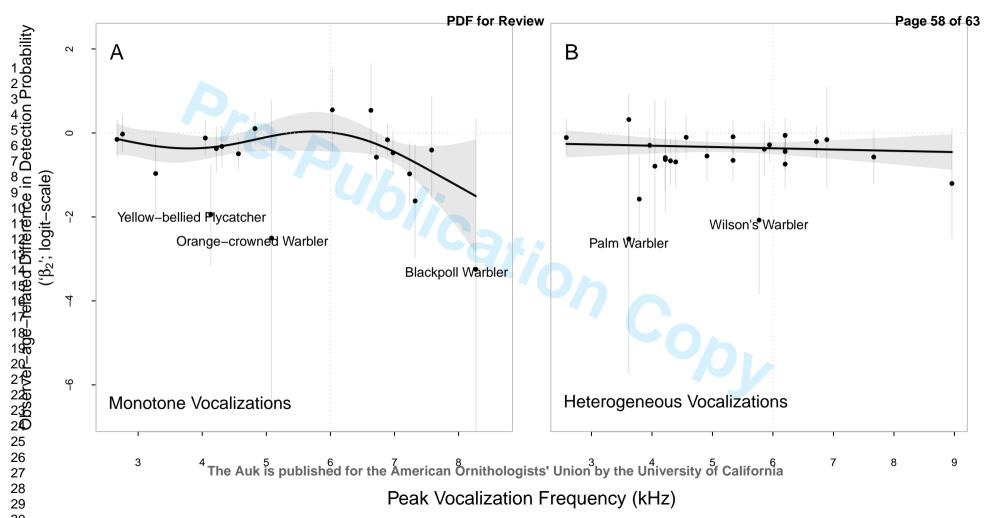911 that observer for the modeled species.

912 To properly recognize the changes in BBS counts predicted by the smooth function

913 $f_1()$ in these models (Equation 8), we did not simply extract its values for the modeled

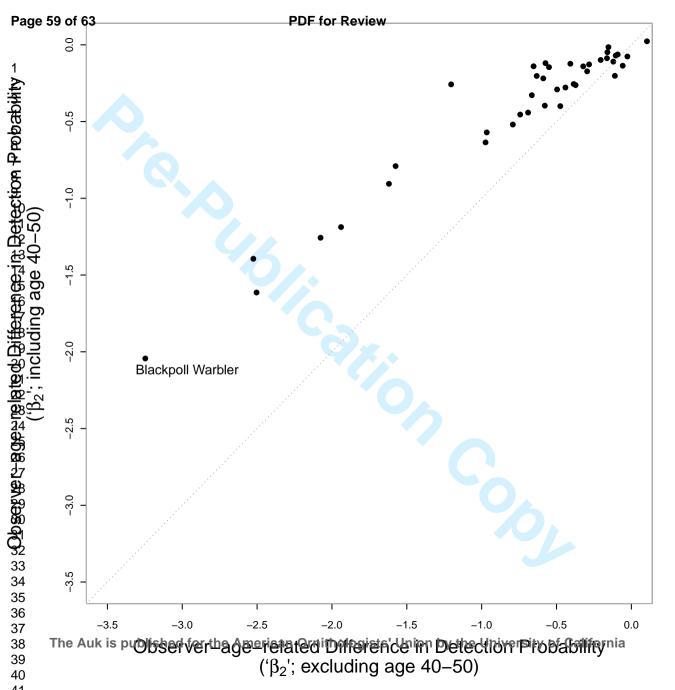914 range of observer ages, since this approach would ignore the uncertainty among the separate

915 population-related smooth terms (estimated for each stratum; $f_2(l)_j$). Instead, working

916 on the scale of the response variable, we defined species- and observer age-specific count

917 predictions as the average of predictions for each relevant physiographic stratum. Calendar

918 years were fixed at the midpoint of surveyed dates during predictions. We inferred the

919 standard error about these averaged predictions, $\bar{\sigma}_{kl}$, as the square root of the mean of their
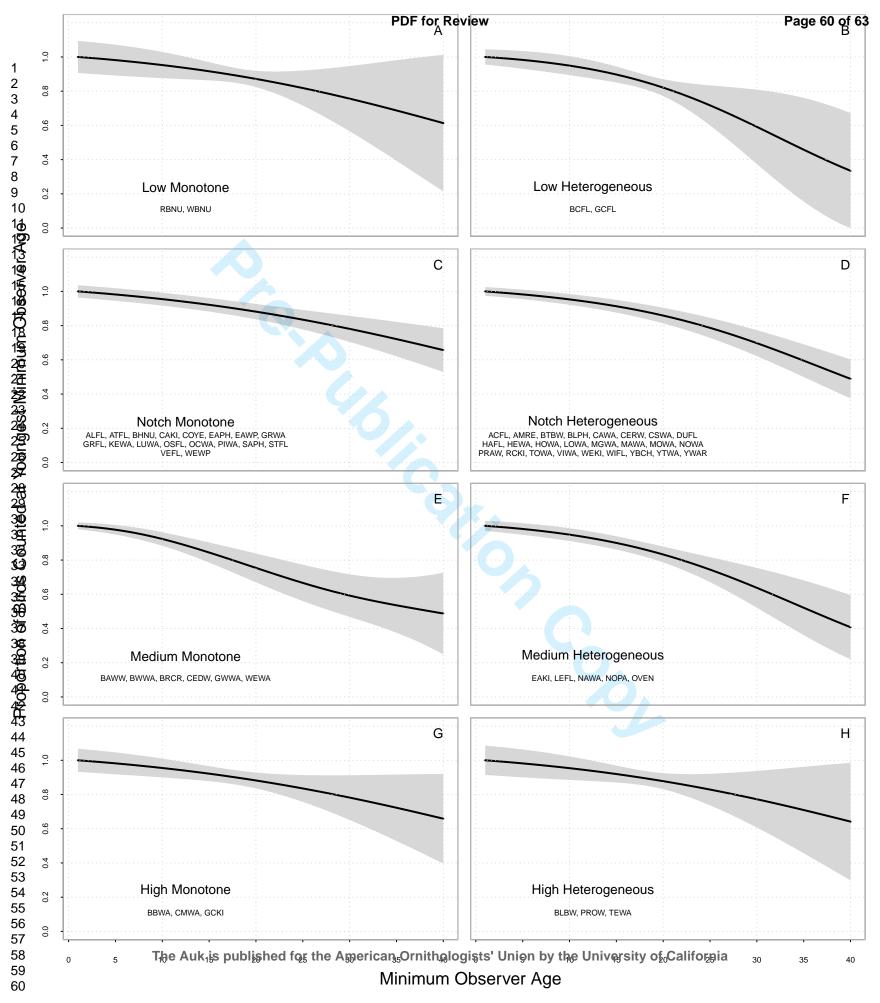
920 variances.

921   We then built an 'aggregating' GAMM which generalized the predicted changes in

922 BBS counts for each species with increasing observer age (produced above) among each of

923 eight vocalization frequency groups (e.g. 'high monotone', 'notch heterogeneous'; discussed

924 in Methods). In addition to generalizing the patterns of age-related count changes among

925 species, this approach also ensured that each species contributed the same number of datapoints

926 to the overall model. To convert the data to a common scale among all species, we used

927 proportions of each species' maximum count as the (binomial) dependent variable in this
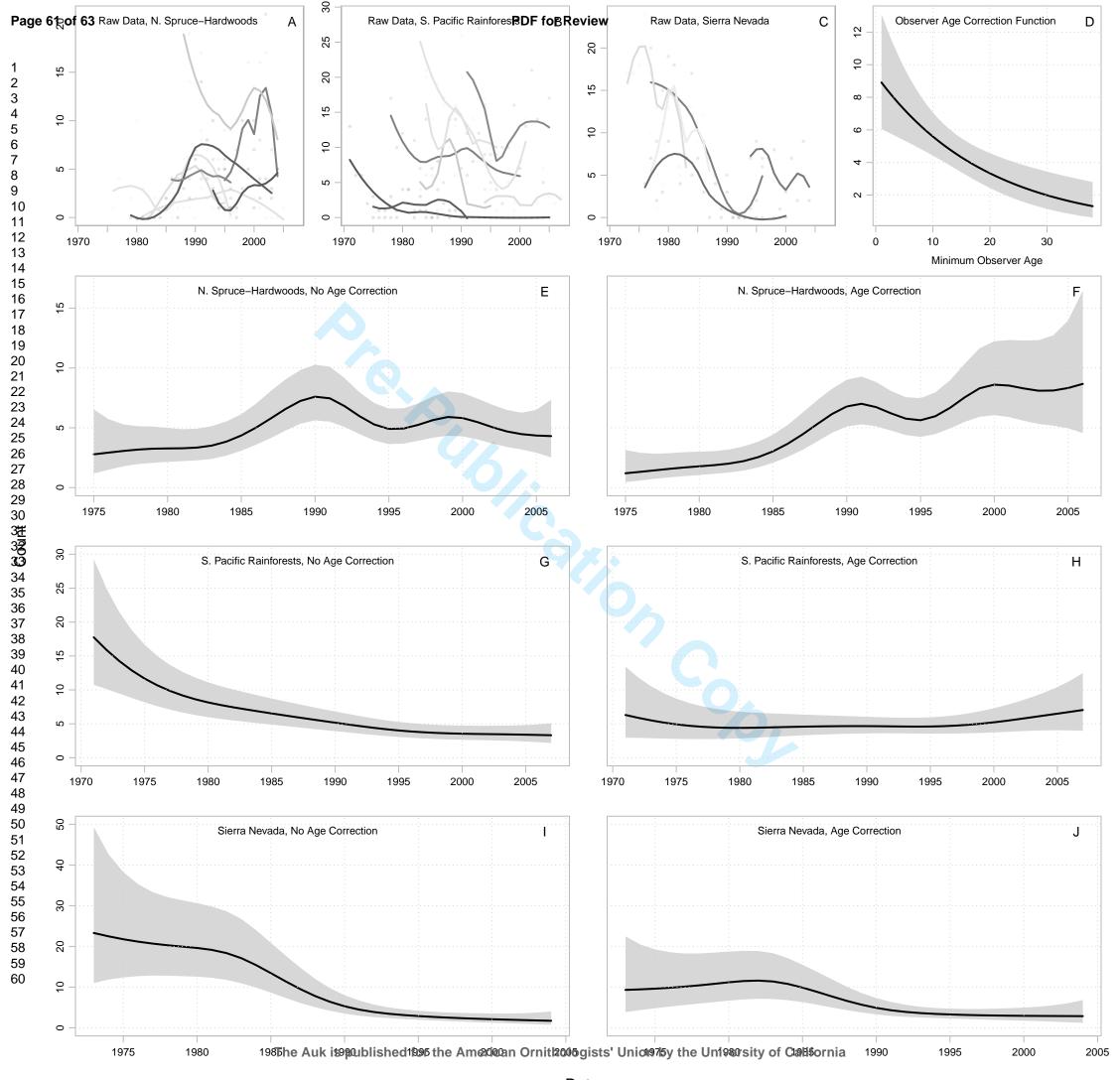
928 model.

929   Similar to the single-species models (equation 8), the aggregating GAMM used thin-plate

930 regression spline smooth functions on observer age for each vocalization group, along with

931 mean-zero, normally-distributed random intercepts for species. Each datapoint was weighted

932 according to the inverse of its predicted coefficient of variation (i.e. $\frac{\hat{\mu}}{\hat{\sigma}}$). To provide a

933 more useful interpretation of the species-independent changes in BBS counts with increasing

934 observer age, final model predictions were then linearly rescaled relative to the values at

935 observer-age 1 for each vocalization group. As in the detection probability analysis, we

936 again validated the performance of our statistical approach by modeling simulated datasets

937 of known population trajectories (see electronic supplement for example code).

Date