

Duck, duck, goose: Benchmark bird surveys help quantify counting errors and bias in a citizen-science database

W. Douglas Robinson^{1*}, Tyler A. Hallman¹, Rebecca A. Hutchinson^{1,2}

¹Oak Creek Lab of Biology, Department of Fisheries and Wildlife, Oregon State University, Corvallis, Oregon, USA

²School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, Oregon, USA

* Correspondence:

W. Douglas Robinson

douglas.robinson@oregonstate.edu

Keywords: biodiversity benchmarks, birder behavior, citizen science, eBird, statistical bias, statistical error, wildlife counts

Abstract

The growth of biodiversity data sets generated by citizen scientists continues to accelerate. The availability of such data has greatly expanded the scale of questions researchers can address. Yet, error, bias, and noise continue to be serious concerns for analysts, particularly when data being contributed to these giant online data sets are difficult to verify. Counts of birds contributed to eBird, the world's largest biodiversity online database, present a potentially useful resource for tracking trends over time and space in species' abundances. We quantified counting errors in a sample of 1406 eBird checklists by comparing numbers contributed by birders (N=246) who visited a popular birding location in Oregon, USA, with numbers generated by a professional ornithologist engaged in a long-term study creating benchmark (reference) measurements of daily waterbird counts. We focused on waterbirds, which are easily visible at this site. We evaluated potential predictors of count differences, including characteristics of contributed checklists, of each species, and of time of day and year. Count differences were biased toward undercounts, with more than 75% of counts being below the daily benchmark value. When only checklists that actually reported a species known to be present were included, median count errors were -29.1% (range: 0 to -42.8 %; N=20 species). Model sets revealed an important influence of each species' reference count, which varied seasonally as waterbird numbers fluctuated, and of percent of species known to be present each day that were included on each checklist. That is, checklists indicating a more thorough survey of the species richness at the site also had, on average, lower counting errors. However, even on checklists with the most thorough species lists, counts were biased low and exceptionally variable in their accuracy. To improve utility of such bird count data, we suggest three strategies to pursue in the future. One is to assess additional options for analytically determining how to select checklists that have the highest probability of including less biased count data, as well as exploring options for correcting bias during the analysis stage. Another is to add options for users to provide additional information that helps analysts choose checklists, such as an option for users to tag checklists where they focused on obtaining accurate counts. We also recommend exploration of opportunities to effectively calibrate citizen-science bird count data by establishing a formalized network of marquis sites where dedicated observers regularly contribute carefully collected benchmark data.

Introduction

42 Contributions of volunteers to scientific databases are increasing as the popularity of citizen science
43 continues to grow (Miller-Rushing et al., 2012; Chandler et al., 2017). Many citizen science projects
44 are open-access and anyone can contribute observations without required training in best data
45 collection practices (Cohn, 2008). eBird is an open online database with more than 560,000 users
46 (eBirders) contributing millions of bird observations annually via checklists (Sullivan et al., 2009).
47 Each checklist contains a list of bird species identified on a particular date and, ideally, counts of
48 each species, as well as information on location visited, basic protocol used while birding (traveling,
49 staying stationary, etc.), and duration of effort (Wood et al., 2011). The huge spatial extent of
50 presence-absence data in eBird has facilitated efforts to model species distributions across continental
51 and global spatial scales once data have been filtered to exclude potentially problematic checklists
52 (Fink et al., 2013). The degree to which the count data may reliably inform scientific and
53 management objectives remains unclear.

54 Although efforts to quantify issues associated with bird species detection have been studied and
55 continue to be developed, both in citizen science databases and in structured scientific surveys
56 (Buckland et al., 2008; Hutto, 2016; Walker and Taylor, 2017), less is known about potential
57 counting errors and biases leading to noisy data. Counting birds is difficult, even by the most
58 proficient observers (Robbins and Stallcup, 1981; Robinson et al., 2018). Methods to account for
59 detection issues in bird counting studies continue to expand with development of new data collection
60 and analytical methods (Buckland et al., 2008; Barker et al., 2018). Nearly all the methods, however,
61 require a sophisticated sampling protocol that would exclude most volunteer birder contributions and
62 therefore limit the advantages of gathering data at massive geographic scales. Yet, the potential
63 windfall from large quantities of data can quickly be eroded if a lack of structured protocols leads to
64 data quality concerns (Kelling et al., 2019). Given that abundance is one of the fundamental
65 influences on population dynamics, functional roles in ecosystems, and even extinction risk (Brown,
66 1984), a better understanding of the potential value of count data contributed to massive online
67 databases by untrained volunteers is needed (Greenwood, 2007). For example, species count errors in
68 eBird data could limit our abilities to observe important abundance trends (Horns et al., 2018).
69 Effective processes for evaluating and handling such errors need further development, owing to the
70 potentially huge value of tracking population changes at a continental scale during this era of rapid
71 environmental change (Bird et al., 2014; Fink et al., 2020).

72 Among the primary concerns are errors, bias and noise. Errors, for our purposes here, are differences
73 in counts between a reference (benchmark) value and values included in eBird checklists for the same
74 species on the same date. Errors are comprised of both bias and noise. Bias is the tendency for the
75 errors to be consistently higher or lower than the reference value. Noise is the additional random
76 counting error that increases variance of the counts. All three impede efforts to determine true count
77 values, and are challenges common to many branches of biology (West, 1999; Guillery, 2002). We
78 acknowledge that labeling such count differences as errors assumes the benchmark values have less
79 error and doing so risks offending some eBird contributors. Given that reliable benchmarks are
80 achieved by consistent application of best counting practices, we do consider deviations from the
81 benchmark values here to be errors, not simply variation among observers. To acknowledge that
82 there are sources of error in all measurements, however, we often refer to such deviations as count
83 differences. We consider the terms ‘error’ and ‘count differences’ to be synonymous.

84 Robust comparisons of count differences are improved when data are collected in situations where
85 detectability challenges are expected to be low. Such situations are rare but uniquely valuable. We
86 used an extensive data set focused on benchmarking the richness and abundances of birds at a water
87 treatment site in Oregon, USA. We compared count data gathered by a professional ornithologist

88 focused specifically on creating an accurate benchmark measurement of daily fluctuations in
89 waterbird counts with counts submitted by birders to eBird. We quantified the magnitude and
90 directionality of count differences. Our data span 10 years and include 1406 eBird checklists
91 contributed by 246 observers, as well as 2038 checklists in the benchmark data. The site is well
92 suited for rigorous comparisons because all waterbirds are in the open, largely tolerant of human
93 activity, and so provide a best-case scenario for detection, identification, and counting of birds. No
94 adjustments for detectability or availability issues should be needed because all parts of the ponds are
95 visible. Thus, discrepancies in counts between a professional observer focused on obtaining accurate
96 numbers and data reported to eBird should be attributable to counting errors instead of availability
97 and detectability issues. While there could be very minor detectability issues, like some diving
98 waterbirds being under water briefly, the vast majority of error in this setting should be attributable to
99 counting error.

100 We first quantified count differences then sought to understand potential factors explaining the
101 magnitude and directionality of count differences. We hypothesized that counting errors would be
102 influenced by traits associated with the species being counted, with an index of observer experience
103 (percent of species detected), and with seasonal changes in numbers of birds present. For example,
104 we expected count differences might be slightly greater for diving ducks, which are sometimes
105 briefly under water while foraging, and lower for dabbling species, which sit in the open
106 continuously. We expected smaller count differences in checklists that included a higher proportion
107 of the species present each day. We also hypothesized that count differences would be greater when
108 overall total number of waterbirds present was high, potentially causing observers to be overwhelmed
109 and therefore more prone to counting errors. Finally, we explored the possibility that, even if count
110 data were biased on individual checklists, the waterbird community might be adequately
111 characterized as a whole by combining count data from multiple observers and checklists. We
112 conclude by proposing additional approaches that may reveal the extent to which citizen-science bird
113 count data may be used to estimate abundances reliably.

114 **Methods**

115 **Study Area**

116 Bird count data were gathered from 2010 to 2019 at the Philomath Wastewater Treatment facility, in
117 Philomath, Oregon USA. The site contained two 35-ha ponds until 2011 when two additional 35-ha
118 ponds were added. Each pond is rectangular and enclosed by a berm with a single-lane road. Birders
119 circumnavigate the ponds typically by vehicle, rarely by walking or bicycling; WDR drove.
120 Vegetation does not obscure the view at any pond. All shores are covered by large rocks (riprap).
121 Birders circle all four ponds during a visit, very rarely restricting visits to fewer ponds. We found that
122 the distribution of visit durations was unimodal (median = 60 min; Median Average Deviation
123 (MAD) = 37; skew=1.161; N=1646 checklists) suggesting that birders use similar methods while at
124 the ponds.

125

126 **Study species**

127

128 We included 20 species we refer to as “waterbirds,” species that swim in the open while on the ponds
129 and should be easily seen (Table 1). The species are primarily ducks and geese, but also include

130 grebes, American Coot (*Fulica americana*), and gulls. These are species birders identify by sight, not
131 by sound. We excluded species that occurred primarily as fly-overs, such as Cackling Goose (*Branta*
132 *hutchinsi*), species whose counts rarely exceeded two per day, and species whose numbers varied
133 strongly within a day. The number of waterbirds present at the site varied seasonally from a few
134 dozen during mid-summer (June) to five thousand or more during fall migration (October-
135 November).

136

137 Benchmark counts

138

139 All birds of all species were counted during each site visit by WDR. We call these our benchmark
140 counts (R^*) and they serve as the reference values against which all other count data are compared.
141 Waterbird counts were made to plus or minus one individual except for Northern Shoveler (*Spatula*
142 *clypeata*), which were plus or minus 10 because they forage in constantly moving dense aggregations
143 rendering more precise counts problematic, and Bufflehead (*Bucephala albeola*), which were counted
144 to plus or minus 5 because they dive so frequently while foraging in the early morning period
145 surveyed by WDR that more accurate counts were difficult. Counts were tallied separately for each
146 pond then aggregated later. In the time frame of these counts, movements between ponds were
147 normally minimal. Duration of counting time was recorded separately for each pond. To reduce
148 possible use of WDR's count data by eBirders who wanted to post numbers but may not have
149 counted on their own, we used three steps to minimize copying of data. First, we imposed a time lag
150 of one to four weeks between dates of counting and of uploading to eBird the WDR data. Second, we
151 hid all of WDR's checklists from the eBird public display in Recent Visits and, third, we posted only
152 the pond-specific data, not the aggregated data. We used aggregated counts from the first visit each
153 day as R^* for comparison with counts reported on eBird checklists.

154 On some days ($N=84$), WDR counted birds more than once. These second-visit data, which we call
155 Ref2 counts, were also complete counts of the study species and averaged 13% shorter in duration,
156 yet counts were generally similar. They were used to characterize within-day variability in numbers
157 but provide a conservative estimate of that variability because they were largely conducted on days
158 with exceptional levels of migratory movements. Thus, they estimate a probable upper bound on the
159 expected amount of within-day variability in waterbird numbers (averaging 0 to -8%). We also used
160 these Ref2 data to evaluate of time-of-day effects when comparing WDR counts with data from the
161 ten observers contributing the most study site data to eBird, because eBirders tended to count birds
162 later in the day than did WDR. The times of day eBird checklists were initiated as well as the
163 difference in start times of eBird and benchmark checklists were unimportant in predicting percent
164 error in our across-species and species-specific model sets. Therefore, we concluded that
165 comparisons of count differences between R^* and eBird checklists were appropriate and that possible
166 time-of-day effects could be ignored.

167

168 eBird checklists

169

170 We downloaded eBird checklists from the Philomath Sewage Ponds eBird hotspot as well as eBirder
171 personal locations within 1 km from 2010 to 2019. Only data obviously restricted to the ponds were
172 included. No other waterbird sites are present within 4 km of the site. Most eBirders used the pre-
173 established hotspot as the checklist location but some created new personal locations each time. We
174 included eBird checklists following the stationary, traveling, and area protocols. We removed
175 checklists with greater than ten observers or durations of over five hours. We included only complete
176 checklists with all birds reported and removed any checklists where observers reported no waterbirds.
177 From each complete eBird checklist, we collected data on date, start time, observer, duration of
178 count, identity of waterbird species reported (to allow calculation of percent richness; see below), and
179 count data for our twenty focal species. When species were recorded as present but not counted (X
180 noted instead of a number), those data were excluded because no count difference could be
181 calculated.

182

183 Comparisons of count data

184

185 We restricted our comparisons to dates where WDR counted birds and at least one eBird checklist
186 was contributed on the same day (N=767 dates). Our questions were about counting differences and
187 not detection rates of rare species, so we further restricted our comparisons to counts of greater than
188 three for each species detected on WDR's first visit (R^*). We calculated the *Count Difference* for
189 each species by subtracting R^* from eBird counts on each checklist. Count differences were positive
190 when eBird checklists reported higher numbers than R^* or negative when eBird checklists reported
191 fewer birds than R^* . Numeric values of count differences spanned three orders of magnitude, so we
192 focus on reporting *Percent Error*, which we calculated by converting each difference to a proportion
193 of R^* .

194

195

196 Hypothesized predictors of percent error

197

198 To evaluate factors hypothesized to be associated with percent error, we included variables
199 associated with species, checklists, time of year and observer experience. *Species characteristics*
200 included categorization as dabbler versus diver, degree to which species form dense aggregations,
201 and the degree of sexual dimorphism. *Checklist characteristics* included start time, duration and
202 number of observers. *Time-of-year characteristics* were associated with daily numbers of waterbirds
203 (R^* , Ref2 and their sums for all 20 species) and waterbird species richness present at the study site
204 (measured as the richness detected by the professional [proRichness] as well as the aggregate of
205 species listed in eBird checklists and proRichness). Because *observer experience* at the site might
206 also influence counting accuracy, we compared data from the 10 observers who contributed the most
207 checklists with the R^* and Ref2 benchmark data. Additional details on each variable are explained
208 below.

209

210 Species characteristics

211 To explore patterns of species-specific variability in count data, we created categorical variables for
212 species traits that might impact counts (Table 1). We categorized birds as dabblers versus divers.
213 Dabblers were any species that foraged primarily by swimming on the surface of the water, which
214 included gulls, American Coot, and *Aix*, *Anas*, *Mareca*, and *Spatula* ducks. Divers foraged below
215 water regularly and included scoters, grebes, and *Aythya* and *Bucephala* ducks.

216 We also included an index of spatial aggregation on the ponds. Some species, for example Northern
217 Shoveler, often forage in densely packed groups, creating challenging circumstances to accurately
218 count birds, while other species forage singly or as spatially-distanced groups where enumeration
219 should be much easier. The aggregation index was simply a subjective binary classification (0 for
220 foraging alone or in loose aggregations versus 1 for foraging in aggregations that might render
221 counting difficult) based on our years of experience at the site.

222 The degree of plumage dimorphism and similarity to other species could influence error and bias in
223 counts because of species misidentification. We categorized species as those with weak or no obvious
224 plumage dichromatism during most of the period of time when each species was present (e.g., geese,
225 coots) versus strong dichromatism (males and females distinctly visually different).

226 To evaluate the possibility that species identification of similar species might influence count
227 differences, we used another subjective binary category called “Doppelganger;” 1 indicated the
228 species co-occurred with a similar species whereas 0 indicated the species was unique in appearance
229 and unlikely to be confused with other species. The categorization may vary seasonally, especially in
230 late summer when many waterbirds molt to eclipse plumage. Because total waterbird numbers were
231 low during late summer, we utilized one value for each species.

232 Checklist characteristics

233 Daily start time among eBird checklists was highly variable, covering all daylight hours. The mean
234 start time was 4 hours later than the mean start time for WDR visits. Although we only compared
235 counts conducted on the same day, we wanted to evaluate potential effects of time-of-day and
236 temporal lag between the eBird checklist counts and R^* . To do so, we converted checklist start time
237 to minutes since midnight then calculated the difference in start time between eBird checklists and
238 WDR first visits.

239 Because our Ref2 counts occurred later in the day when more eBird checklists were initiated, we
240 included Ref2 as an “additional observer” in some comparisons to provide an important check on
241 within-day variability in counts as a possible explanation for count differences between R^* and eBird
242 checklists. Because Ref2 counts were generated on days with high levels of migratory movement, we
243 consider the count differences between R^* and Ref2 to represent an upper bound on expected levels
244 of within-day variability in waterbird numbers.

245 Additional factors associated with each checklist could influence count differences. We reasoned that
246 duration of time spent at the site should be positively related to count accuracy. All complete eBird
247 checklists are required to have a measurement of event duration.

248 Number of observers might also influence counting accuracy, so we included the reported number of
249 observers for each eBird checklist. The R^* and Ref2 counts were made when WDR was alone more
250 than 99% of all dates.

251

252 Time-of-year characteristics

253 Date influences the number of species present as well as the abundances of each species. Both
254 richness and abundance could influence counting accuracy so we included day of year in our models.
255 Because we hypothesized that total number of all waterbirds combined may influence counting
256 accuracy, we included R^* counts of all 20 study species and the combined daily total of all waterbirds
257 in our model sets. In that way, we established the baseline numbers of waterbirds known to be
258 present as a function of date. In calculating total waterbird abundance, we used data limited to the 20
259 study species and excluded a subset of species known to have high daily variability in counts, such as
260 geese, which occurred primarily as fly-overs. The other species excluded from our focal group of 20
261 species were numerically rare. Further, to determine if percent error was influenced by the number of
262 each particular species as opposed to overall waterbird abundance, we included R^* of each relevant
263 species in our model sets.

264 We hypothesized overall waterbird species richness present at the site on a given date may influence
265 counting accuracy. A higher number of species to identify could reduce focus for achieving accurate
266 counts, particularly for the more regularly-occurring and common species (e.g., Mallards, Northern
267 Shovelers). Therefore, we included in our models the total waterbird richness detected by WDR each
268 day. Our analyses indicated that richness observed by WDR and total waterbird richness detected by
269 all eBird contributors were highly correlated. We calculated daily *Percent Richness* based on the 35
270 possible waterbird species at the site and included that richness in our models (see Supplemental Text
271 for a list of species). The other 15 species that formed our set of 35 waterbird species included: Snow
272 Goose (*Anser caerulescens*), Greater White-fronted Goose (*Anser albifrons*), Cackling Goose
273 (*Branta hutchinsii*), Canada Goose (*Branta canadensis*), Blue-winged Teal (*Spatula discors*),
274 Eurasian Wigeon (*Mareca penelope*), Redhead (*Aythya americana*), Tufted Duck (*Aythya fuligula*),
275 Greater Scaup (*Aythya marila*), White-winged Scoter (*Melanitta deglandi*), Black Scoter (*Melanitta*
276 *americana*), Long-tailed Duck (*Clangula hyemalis*), Common Goldeneye (*Bucephala clangula*),
277 Barrow's Goldeneye (*Bucephala islandica*), and Common Merganser (*Mergus merganser*).

278

279 Observer experience

280

281 Observer experience at the site could also be influential, so we compared percent error in counts from
282 the ten observers contributing the most eBird checklists at our study site with the R^* and Ref2 counts.

283

284 Data analyses

285

286 We used the “lmer” package in R (R Core Team, 2020) to run mixed-effects models. Our
287 overarching goal was to identify factors informative for explaining variation in *Percent Error*, our
288 dependent variable in all models. We included observer ID and species as random effects to account
289 for observer- and species-specific error when appropriate. We included four categorical species

290 characteristics as fixed effects in our model sets: Dabbler or Diver; Sexually Dichromatic or not;
291 Doppelganger or not; and Aggregated or not. Five checklist-related characteristics were included as
292 fixed effects: start time (minutes since midnight), difference in start time between WDR's first count
293 of a day and each eBird checklist, duration (minutes), number of observers, and day of year. Four
294 fixed-effects related to time-of-year were also included: R^* (WDR's reference count of each species,
295 which varied seasonally), waterbird abundance (aggregated across all species), total waterbird species
296 richness and percent richness, our index of observer skill at species identification. We included
297 models with the quadratic effects of species-specific abundance, waterbird abundance, waterbird
298 richness, duration, number of observers, day of year, and percent richness to examine potential
299 nonlinear shapes of their effects.

300 Before running mixed effects models, we scaled and centered all numeric variables. We assessed
301 model performance through BIC and propagated best-performing shapes for each variable to multi-
302 variable models. We used a forward stepwise approach and added additional potentially influential
303 variables to the best-performing model until a stable (i.e., model remained the top model after the
304 inclusion of additional variables) top-performing BIC model was identified.

305 Although *count difference* was normally distributed, *percent error* was not. Non-detections of species
306 that were detected by WDR (eBird counts of zero) equal negative 100 *percent error*. Non-detections
307 caused a bimodal distribution of *percent error* with a second peak at negative 100 percent. We
308 removed non-detections to create a unimodal distribution of percent error. When non-detections were
309 removed, *percent error* was heavily right-skewed due to the high number of negative *percent errors*
310 and the few very large positive *percent errors*. To adjust skew, we added a constant to make all
311 values positive and log (base 10) transformed percent error. In addition to adjusting skew, removal of
312 non-detections improved the focus of our analyses on count errors, reducing chances that inclusion of
313 zero counts of species might actually be species detection or identification problems instead of
314 counting errors. Our restriction of counting error analyses to species detected in numbers of 3 or
315 greater probably limited most effects of zero counts. In this paper we focus on analyses of data
316 excluding non-detections but report some analyses in supplemental materials to show the effects of
317 including non-detections (zero counts) on results. It is possible that an unknown number of zero
318 counts were a result of reporting errors (data entry mistakes), but we assume this type of error is
319 relatively rare.

320

321 Species-specific model sets

322

323 To understand the (in)consistency of variables influencing species-specific percent error, we ran
324 standardized linear model sets of the effects of the explanatory variables described above on
325 *transformed percent error* for each species. As above, we included models with quadratic effects of
326 species abundance, waterbird abundance, waterbird richness, duration, number of observers, day of
327 year, and percent richness. As each model set was species-specific, we excluded variables of species
328 characteristics from these model sets. We included observer ID as an explanatory variable to examine
329 its comparative influence. In these standardized model sets, we included separate models of the main
330 effect of each variable and propagated the best shape for each variable into more complex models.
331 Since start time and difference in start time were highly correlated, we use the top-performing of the
332 two in subsequent models. We used a forward step-wise approach to determine the top-performing

333 model of checklist covariates. We then ran models with pairs of all non-checklist explanatory
334 variables with and without the variables in the top checklist covariate model. We used BIC to
335 compare model performance and select top models.

336

337 Non-metric Multidimensional Scaling (NMDS)

338

339 To compare the overall communities described in eBird checklists, we conducted ordination in
340 species space with NMDS on count data. We grouped checklists by observers to simplify the
341 analysis. To visualize differences in community characterization, we chose to contrast January and
342 October because January represents a time of year when waterbird migration is minimal and so daily
343 numbers are relatively stable, whereas migration is at its peak during October, so richness is high and
344 volatility in numbers can be high. To evaluate how characterization of waterbird abundance at these
345 times varied with respect to eBirder checklists, we first removed all checklists that included an “X”
346 for the count of any of our 20 study species. We then calculated the mean and median values of
347 species counts across checklists for each observer during each month. To evaluate the idea that group
348 collective contributions of multiple eBird checklists might characterize the waterbird community
349 more similarly to R^* , we calculated mean counts of species across observers in January and October
350 to create combined count values, which we call the Borg number (\bar{B}). We similarly aggregated
351 WDR’s first-visit species counts as a Reference community. To ensure that our \bar{B} NMDS positions in
352 species space were not driven overwhelmingly by an eBirder with the largest number of checklists,
353 we reran the NMDS without checklists from the top-contributing observer included in \bar{B} . We used
354 two dimensions and a maximum of 20 iterations to run NMDS with the “vegan” package in R
355 (version 3.6.1).

356

357

358 **Results**

359

360 We compared benchmark counts of waterbirds from WDR (R^*) and at least one eBirder on 672 dates,
361 representing a total of 1406 comparisons (checklists). eBird checklist contributions varied seasonally
362 with lows during winter and summer and highs during migration periods (Supplemental Figure 1).
363 Our analyses included 246 different eBirders who contributed from 1 to 321 checklists.

364

365 **Percent error**

366 Across all twenty species, 76 percent of all counts fell short of R^* (Figure 1, Supplemental Figure 2),
367 indicating that count data in eBird checklists regularly contained apparent counting errors. eBird
368 checklists with species non-detections excluded (that is, no counts of zero included, even if the
369 species was known to be present that day) had counts below R^* values by a median of 29.1% but
370 errors were quite variable across species (Figure 1a), with median absolute deviations of *percent*

371 *error* averaging 44.6% (Supplemental Table 1). At the extremes, count differences across waterbird
372 species ranged from negative 99% for severe under-counts to more than 3788% too large. In real
373 numbers, counting differences ranged from being too low by 1443 to too high by 1048 (both for
374 Northern Shoveler; Figure 1b). Median percent error was negative, indicative of undercounting, for
375 all waterbird species except the uncommon Surf Scoter (0%; R^* was always less than 11).

376 Percent error, when averaged across species and all observers, was fairly consistent at 30% when
377 counts were 30 or greater. Below thirty, counts were more accurate, being closest to zero error when
378 counts were of 8-10 birds (Figure 2A). Percent error was related to the percent richness (our index of
379 observer skill where higher percentages indicated an observer included more of the species known to
380 be present that day on their checklists) in a curvilinear fashion. Checklists including the lowest
381 richness tended to overcount (Figure 2B). Those including 50% of the expected species undercounted
382 by 50% on average, while checklists including 90% or more of the species reported on R^* checklists
383 averaged errors of 15% or less in count.

384

385 BIC Top models

386 In our multi-species mixed-effects model set, our top model garnered 70 percent of the model weight
387 and was over four BIC from the next most competitive model (Table 2). Our BIC top model
388 indicated that a quadratic effect of R^* and a linear effect of percent richness best explained variation
389 in percent error.

390 Seasonality in bird numbers was also captured when the second-order R^* was included as the most
391 informative variable predicting *percent error*. Numbers of all species varied considerably across each
392 year (Figure 3). Likewise, total waterbird abundance varied several-fold from its nadir in June to a
393 maximum in October and November (Supplemental Figure 3). Yet, total waterbird abundance was
394 rarely an informative variable in our model sets. Only in counts of American Coot did it appear in the
395 most parsimonious models (in combination with percent richness). In California Gull, waterbird
396 abundance appeared as an informative variable but only in a weakly competitive model (19% of the
397 model weight).

398 Within the species-specific model sets, the combination of R^* and percent richness carried most of
399 the model weight (mean=0.83, SD=0.18) in 13 of our 18 non-gull species (Supplemental Table 2).
400 For gulls, top models struggled to outcompete the null. Altogether, R^* and/or percent richness were
401 in the top model sets for 17 of 18 non-gull waterbirds.

402

403 Associations with bird characteristics

404 Within our full model, bird characteristics were rarely influential on percent error (Table 2).
405 Similarly, species-specific models rarely discovered bird traits to be informative variables
406 (Supplemental Table 2).

407

408 Observer effects

409 Our models often identified percent richness as an influential variable on percent error, so we related
410 percent richness to percent error as means across all checklists contributed by each observer (Figure
411 4a). The two were positively related, yet only six of the 246 observers averaged *percent errors* of less
412 than 10%. The range in percent error for observers detecting 90% or more of waterbird species was
413 actually greater than the range for observers who detected less than 60% of species, indicating that
414 percent error alone is an unreliable predictor of counting accuracy. The relationship was not
415 necessarily driven by site experience because four of the six observers with the most accurate counts
416 were contributing very few checklists (Figure 4b).

417 We then selected checklists from the ten observers who contributed the most. Those checklists also
418 showed evidence of undercounting. In nearly all 20 species, percent error was 10 to 60% greater than
419 even the Ref2 counts (Figure 5). Percent error was highly variable across species. In some species,
420 such as American Coot, three of the 10 observers reported counts averaging very near the Ref2
421 counts, whereas in others, such as Pied-billed Grebe, all observers undercounted by at least an
422 average of 20%. Again, percent error was highly variable in all species even when median percent
423 error did not deviate far from zero.

424

425 Community visualization

426 We visualized characterization of the richness and abundance of the daily waterbird community with
427 NMDS through ordination of checklists (grouped by observer) in species space. Observers
428 characterizing the community and its species abundance patterns similarly to R^* fell nearer to R^*
429 whereas those positioned increasingly further from R^* described the community in increasingly
430 dissimilar details. In both January (Figure 6a) and October (Figure 6b) high inter-observer variability
431 in how their checklists characterized the waterbird community led to a general lack of clustering near
432 R^* . In both months, observers reporting more species, contributing more checklists, and surveying
433 for more time tended to group nearer R^* . The collective average, \bar{B} , was nearer R^* than any
434 individual observer during January but one observer was closely positioned near \bar{B} during October.
435 Removal of checklists from the observer contributing the most data had minimal effects on results.

436

437

438 Discussion

439 Benchmark data are often designed to understand temporal change in biodiversity (Curtis and
440 Robinson, 2015; Curtis et al., 2016; Robinson and Curtis, 2020). Here, we show that they can also be
441 used to establish standards that aid in quantification of potential errors in citizen-science data.
442 Through comparisons with such a standard, we discovered that bird count data contributed to eBird
443 from our study site were consistently biased toward undercounting. Counts averaged approximately
444 30% too low whenever benchmark counts were of 30 or more birds. Importantly, however, errors
445 exhibited high variability across species and observers. Benchmark data like ours can subsequently
446 inform decisions regarding what subsets of data should be selected to most rigorously address
447 particular scientific questions or management decisions, analogous to how checklist calibration
448 indices help researchers choose suitable eBird checklists based on site- and time-specific
449 expectations of species richness (Yu et al., 2010; Kelling et al., 2015; Johnston et al., 2018). Yet,

450 situations in which such informative standards may be developed and compared appear to be rare
451 currently.

452 Our study site presented a unique opportunity to compare bird count data contributed to a citizen
453 science database (eBird) with benchmark reference data collected by a professional observer focused
454 on generating accurate daily counts. Characteristics of the site, where all birds were in the open and
455 identified by sight, minimized issues of availability and therefore the need for detectability
456 adjustments to compare counts. Data were contributed by 246 observers and included 676 dates
457 across 10 years, providing an unusual opportunity to explore patterns and potential sources of error.
458 Although the extent to which our results may be generalized to other sites remains unclear given the
459 rarity of opportunities like this one, the situation probably represents a best-case scenario given that
460 birds were in the open and easy to observe. Despite the advantages, count differences in 20 species of
461 waterbird were highly variable across the calendar year, species, and observer. Coefficients of
462 variation were high, averaging 6.6 across the 20 species and ranging from 1 to 35.6. For comparison,
463 in an experimental study of observer counting errors of singing birds, which should have been much
464 harder to detect and identify but had a lower range of abundances than our waterbird community,
465 coefficients of variation averaged 0.1 (Bart, 1985).

466 Our quantification of counting error is actually conservative because we excluded counts of zero on
467 eBird checklists, even for species known to be present. We did so to minimize the potential confound
468 of misidentifications and reporting errors (failing to enter a count for a species that was actually
469 observed) from our analysis of counting errors. Yet, it is possible that some fraction of 100%
470 undercounts were indeed counting errors in the sense that the species was one that observers were
471 knowledgeable enough to identify but failed to count or report. The median percent error across the
472 20 species was -48.6 plus or minus 50.9% (MAD) when zero counts were included versus -29.1 plus
473 or minus 44.6% when zero counts were excluded. Inclusion of zero counts, therefore, has a large
474 influence on the median, but percent errors were highly variable regardless.

475 Our top overall mixed-effects model carried nearly 70% of the model weight and contained only two
476 variables. The species-specific R^* count as a quadratic, which captured the seasonality in numbers
477 present at the site, was the most informative variable when combined with a linear effect of percent
478 richness. The inclusion of R^* indicates that eBird count data were related to the benchmark numbers
479 but that other factors were also influential. Checklists with a more complete list of the species known
480 to be present each day had lower counting errors. Yet, checklists including 100% of expected species
481 still undercounted by an average of 15%. Count differences on checklists from the ten observers who
482 most often visited the site were still exhibiting undercounts even compared to the Ref2 values, which
483 were benchmark counts made later each day during weeks with high levels of migratory movements.

484 We documented strong directional bias toward undercounts and also a smaller percentage of large
485 overcounts, leading to inconsistent patterns in count differences across species. Our comparisons
486 revealed that undercounting was pervasive, yet very large numbers of a species being present
487 sometimes led to severe overcounting as well. Interestingly, the influence of number of birds
488 appeared to be species-specific. The total number of waterbirds of all species present on a given day
489 was not an influential variable in our overall model explaining percent error, except for one species,
490 American Coot. This pattern suggests that count differences were unlikely to have been caused by
491 observers being overwhelmed by the total number of birds to observe, identify and count. Instead, it
492 appears that each species presented different challenges to observers. Given that our models rarely
493 identified species' traits as being informative, it remains unclear what species-specific factors are
494 responsible.

495 The degree of variability across species in count differences should influence potential decisions
496 regarding use of eBird count data. Our analyses clearly reveal that off-the-shelf acceptance of count
497 data for assessments of absolute abundance should be done with great care and thoughtfulness. In
498 addition, if researchers wish to avoid focus on absolute abundance by instead evaluating relative
499 abundance, our results suggest further caution is warranted. We found great interspecific variability
500 in count differences. That is, although bias was nearly uniformly directional toward undercounting,
501 the magnitude of undercounts varied substantially across species indicating that processes generating
502 errors are inequivalent across species. Therefore, judging differences in one species' abundance
503 relative to others requires careful thought. If explorations of relative abundance are focused on
504 within-species changes across sites, care is also warranted because we found substantial differences
505 among observers in count accuracy. If different sites have different observers, then error/bias
506 processes will be expected to be different as well. Effective use of relative abundance data depends
507 on assumptions of consistent errors across species and sites, which appears to be largely untrue in our
508 data. Further exploration of techniques to determine the degree to which assumptions of similar
509 counting errors across species might be relaxed to preserve the utility of relative abundance analyses
510 are warranted. The use of abundance categories could be explored to maximize the information
511 content gleaned from count data.

512 What role might species misidentifications have played in counting errors? Count differences were
513 regularly so large that we conclude species misidentification was unlikely to be an important factor.
514 Probably the most challenging identifications involved female or eclipse-plumaged ducks, which
515 observers might ignore and exclude from checklists if identification is uncertain. We consider such
516 omissions to be unlikely for at least three reasons. First, degree of dichromatism was uninformative
517 in our models explaining percent error. Second, assuming that females represent approximately half
518 of each species present during most months of a year, count differences might be expected to average
519 50% if males were counted accurately but females were not. Instead, percent error varied widely
520 across species. Finally, count differences of monochromatic versus dichromatic species were not
521 obviously different. However, it is possible that observers were more accurate for some species than
522 others because of paying greater attention to unusual or favorite species (Schuetz and Johnston,
523 2019). At our site, most charismatic species of great interest to birders are rarities and so were not
524 included in our analyses. Counts of Surf Scoter, a species that occurs during a narrow window of
525 time in fall, were generally accurate, but we cannot attribute the accuracy to celebrity alone given its
526 occurrence in such small numbers.

527 Based on our analyses of count differences at this site, it appears that count data on eBird checklists
528 from similar situations should be used with great care and thoughtfulness. Aside from a
529 predominantly directional bias toward undercounts, we found few consistent species-specific patterns
530 in percent error. Errors differed in magnitude across species, observers, and time of year. Therefore,
531 development of some type of calibration effort, where checklist numbers are adjusted to more closely
532 approximate species-specific abundances poses an interesting challenge. The variability in raw count
533 data suggests that tracking trends across time without additional steps to filter data or analytically
534 adjust for noise could be especially problematic. Depending on the particular scientific question of
535 interest, needs for precision might decline, so other analytic approaches could be implemented. For
536 example, if abundances can be binned into categories and approaches such as ordinal or quantile
537 regression used (Ananth and Kleinbaum, 1997; Koenker and Hallock, 2001; Howard et al., 2014),
538 less precisely defined trends over time might be identified. Furthermore, our observation that percent
539 richness, which we assume to be a correlate of observer experience, was often an informative
540 variable, suggests that additional exploration of count calibration approaches for data contributed by
541 the most experienced observers might be informative.

542 If questions about patterns in abundances among species in the waterbird community are of interest,
543 our NMDS ordination results suggest that combining checklists across multiple observers may
544 produce results closer to those generated by professional benchmark data. The vectors in NMDS
545 results may also inform decisions about which criteria to use when filtering data to maximize
546 inclusion of checklists with the greatest value for specific scientific questions. For example, the
547 waterbird community at our site was better characterized by observers who included more species on
548 their checklists, invested more time searching the site each time, and contributed more checklists
549 overall. Although species-specific numbers remained inconsistently related to the R^* counts, the level
550 of general characterization of the entire community was improved. In a detailed comparison of eBird
551 data with structured survey results near Sydney, Australia, overall characterization of the bird
552 communities was similar as well, but the collectively greater effort expended by eBirders resulted in
553 discovery of a greater number of uncommon species (Callaghan et al., 2018).

554 Determining the extent to which results from our site and observers may be generalized more widely
555 will require identification of other sites with benchmark data sets. We also recommend further
556 investigation of approaches for identifying checklists with higher probability of having the most
557 accurate count data. New approaches for categorizing checklists based on expected numbers of
558 species have recently been developed but it remains unclear if these same criteria also apply to bird
559 counting accuracy (Callaghan et al., 2018). Our index of checklist quality was based solely on the
560 percent of species reported on checklists that were also detected that day by the professional
561 observer. Percent richness was regularly in top models, so does have explanatory influence on count
562 differences. Yet, direct comparisons of data from those observers and the R^* and Ref2 numbers still
563 showed substantial differences, primarily of undercounting.

564 If a sufficiently detailed benchmark data set is available, however, adjustments for seasonal
565 fluctuations in numbers of each species could conceivably be implemented. Such calibrations might
566 be conducted more effectively if individual observers exhibited consistency in counting errors, an
567 issue we have not explored here. It is unknown if observers improve their counting skills over time in
568 the same way that observers are expected to improve abilities to detect species or if temporal
569 stochasticity drives counting errors. A goal could be to develop a count calibration metric for each
570 observer so that it can be extended and applied to counts from sites lacking data from a professional
571 observer if those sites are likely to have similar species composition and relative abundances.
572 However, given the high level of variability in count data we quantified across observers, species and
573 time, such calibration metrics may be quite challenging to develop. Complex models such as the
574 Bayesian hierarchical models using Markov chain Monte Carlo approaches like those implemented
575 with Christmas Bird Count data (Link et al., 2006), might be helpful in the absence of additional
576 information on checklist accuracy and reliability. Our community ordination results suggested that
577 combining data across multiple checklists from multiple observers (the group collective effort) might
578 more closely approximate the community characterization than most single contributors did. Further
579 exploration of similar approaches and sensitivities to checklist characteristics could identify
580 necessary checklist quality criteria that must be met prior to use in such analyses. In the end, use of
581 any checklist count data will be influenced strongly by each project's specific objectives (Isaac and
582 Pocock, 2015).

583 We hypothesize that the high variability in species count information on eBird checklists could be
584 influenced by common aspects of birder behavior. Prior to the advent of eBird, most birders, in North
585 America at least, focused their efforts on listing species and watching behavior (Eubanks, Jr. et al.,
586 2004). Intentional counting was done by a small percentage of particularly avid observers, while
587 most others only counted during organized activities such as Christmas Bird Counts (Boxall and

588 McFarlane, 1993). A much smaller percentage contributed count data to scientific projects with
589 structured protocols such as the North American Breeding Bird Survey. eBird has revolutionized the
590 degree of attention birders pay to numbers of birds around them (Wood et al., 2011). It has pushed
591 birders to value data beyond the day's species list. The novelty of this effort to count all birds every
592 time one goes birding, may contribute to the variability in quality of the count data. Contributors are
593 largely untrained about best practices for counting, especially when birds are present in large
594 numbers, flying, or inconspicuous because they are secretive or available only by sound. We
595 encourage development of additional training opportunities for eBird contributors that improve their
596 knowledge of the value of accurate count data as well as their counting skills. Training improves data
597 quality even for professional observers (Kepler and Scott, 1981).

598 An indication on checklists in the eBird database that such training had been accomplished might
599 facilitate selection of checklists by researchers who wish to use count data only from trained
600 observers. Furthermore, the addition of a qualitative categorization of counting accuracy for each
601 checklist, designated by the observer at time of checklist submission to eBird, might be useful.
602 Currently, users may code species using presence-absence information instead of counts or select a
603 checklist protocol (incidental) indicating that not all species detected were included the list. A count
604 accuracy designation could allow observers to rate their own level of confidence in the accuracy of
605 their counts or the level of attention they paid to counting accurately, which could serve as additional
606 criteria by which researchers might choose checklists for their particular scientific question. Given
607 that many contributors may not necessarily participate to contribute data useful for abundance
608 analyses but have a variety of other motivations (Boakes et al., 2016), allowing observers to
609 categorize quickly and easily their personal confidence in their count data would be useful.

610 Finally, exploration of the sources of variation in count data needs additional attention (Dickinson et
611 al., 2010). The potential value of the vast quantities of information from citizen science databases is
612 great. Such data have the potential to be effective at informing conservation and management
613 decisions (McKinley et al., 2017; Young et al., 2019), but a thorough understanding of sources of
614 error should be a priority before their use (Lewandowski and Specht, 2015). An additional strategy
615 that may contribute to refinement of information on count data quality in citizen science databases
616 could be development of a network of sites with trained counters. These marquis sites could be
617 chosen to represent major habitat types where citizen science data are often gathered or where
618 researchers specifically need high-quality information. Creating a network of high-quality benchmark
619 sites would have the added advantage of leaving a legacy of more reliable abundance data for future
620 generations, especially if complete metadata are also preserved.

621

622 **Acknowledgements**

623 WDR and TAH were supported by the Bob and Phyllis Mace Professorship and the College of
624 Agricultural Sciences. RAH was supported by NSF Grant 1910118. We thank the many birders who
625 contributed their data to eBird and the many scientists who created, maintain and continue to improve
626 eBird. Dennis Lewis and Philomath Public Works permitted access to the site, not only to us but to
627 more than 250 birders. We thank our Reconfiguration Grant Group for helpful discussions.

628

629

630 **Literature Cited**

631

632 Ananth, C. V., and Kleinbaum, D. G. (1997). Regression models for ordinal responses: a review of
633 methods and applications. *Int J Epidemiol* 26, 1323–1333. doi:10.1093/ije/26.6.1323.

634 Barker, R. J., Schofield, M. R., Link, W. A., and Sauer, J. R. (2018). On the reliability of N-mixture
635 models for count data. *Biometrics* 74, 369–377. doi:10.1111/biom.12734.

636 Bart, J. (1985). Causes of recording errors in singing bird surveys. *The Wilson Bulletin*, 161–172.

637 Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., et al. (2014).
638 Statistical solutions for error and bias in global citizen science datasets. *Biological*
639 *Conservation* 173, 144–154. doi:10.1016/j.biocon.2013.07.037.

640 Boakes, E. H., Gliozzo, G., Seymour, V., Harvey, M., Smith, C., Roy, D. B., et al. (2016). Patterns of
641 contribution to citizen science biodiversity projects increase understanding of volunteers'
642 recording behaviour. *Scientific reports* 6, 33051.

643 Boxall, P. C., and McFarlane, B. L. (1993). Human Dimensions of Christmas Bird Counts:
644 Implications for Nonconsumptive Wildlife Recreation Programs. *Wildlife Society Bulletin*
645 *(1973-2006)* 21, 390–396.

646 Brown, J. H. (1984). On the Relationship between Abundance and Distribution of Species. *The*
647 *American Naturalist* 124, 255–279. doi:10.1086/284267.

648 Buckland, S. T., Marsden, S. J., and Green, R. E. (2008). Estimating bird abundance: making
649 methods work. *Bird Conservation International* 18, S91–S108.
650 doi:10.1017/S0959270908000294.

651 Callaghan, C. T., Martin, J. M., Major, R. E., and Kingsford, R. T. (2018). Avian monitoring –
652 comparing structured and unstructured citizen science. *Wildl. Res.* 45, 176–184.
653 doi:10.1071/WR17141.

654 Chandler, M., See, L., Copas, K., Bonde, A. M., López, B. C., Danielsen, F., et al. (2017).
655 Contribution of citizen science towards international biodiversity monitoring. *Biological*
656 *Conservation* 213, 280–294.

657 Cohn, J. P. (2008). Citizen Science: Can Volunteers Do Real Research? *BioScience* 58, 192–197.
658 doi:10.1641/B580303.

659 Curtis, J. R., and Robinson, W. D. (2015). Sixty years of change in avian communities of the Pacific
660 Northwest. *PeerJ* 3, e1152. doi:10.7717/peerj.1152.

661 Curtis, J. R., Robinson, W. D., and McCune, B. (2016). Time trumps habitat in the dynamics of an
662 avian community. *Ecosphere* 7, e01575. doi:10.1002/ecs2.1575.

663 Dickinson, J. L., Zuckerberg, B., and Bonter, D. N. (2010). Citizen Science as an Ecological
664 Research Tool: Challenges and Benefits. *Annual Review of Ecology, Evolution, and*
665 *Systematics* 41, 149–172. doi:10.1146/annurev-ecolsys-102209-144636.

- 666 Eubanks, Jr., T. L., Stoll, J. R., and Ditton, R. B. (2004). Understanding the Diversity of Eight Birder
667 Sub-populations: Socio-demographic Characteristics, Motivations, Expenditures and Net
668 Benefits. *Journal of Ecotourism* 3, 151–172. doi:10.1080/14664200508668430.
- 669 Fink, D., Auer, T., Johnston, A., Ruiz-Gutierrez, V., Hochachka, W. M., and Kelling, S. (2020).
670 Modeling avian full annual cycle distribution and population trends with citizen science data.
671 *Ecological Applications* 30, e02056. doi:10.1002/eap.2056.
- 672 Fink, D., Damoulas, T., and Dave, J. (2013). Adaptive Spatio-Temporal Exploratory Models:
673 Hemisphere-wide species distributions from massively crowdsourced eBird data. in *Twenty-*
674 *Seventh AAAI Conference on Artificial Intelligence* Available at:
675 <https://www.aaai.org/ocs/index.php/AAAI/AAAI13/paper/view/6417> [Accessed May 21,
676 2020].
- 677 Greenwood, J. J. D. (2007). Citizens, science and bird conservation. *J Ornithol* 148, 77–124.
678 doi:10.1007/s10336-007-0239-9.
- 679 Guillery, R. W. (2002). On counting and counting errors. *Journal of Comparative Neurology* 447, 1–
680 7. doi:10.1002/cne.10221.
- 681 Horns, J. J., Adler, F. R., and Şekerciöğlü, Ç. H. (2018). Using opportunistic citizen science data to
682 estimate avian population trends. *Biological Conservation* 221, 151–159.
683 doi:10.1016/j.biocon.2018.02.027.
- 684 Howard, C., Stephens, P. A., Pearce-Higgins, J. W., Gregory, R. D., and Willis, S. G. (2014).
685 Improving species distribution models: the value of data on abundance. *Methods in Ecology*
686 *and Evolution* 5, 506–513. doi:10.1111/2041-210X.12184.
- 687 Hutto, R. L. (2016). Should scientists be required to use a model-based solution to adjust for possible
688 distance-based detectability bias? *Ecological Applications* 26, 1287–1294.
689 doi:10.1002/eap.1385.
- 690 Isaac, N. J. B., and Pocock, M. J. O. (2015). Bias and information in biological records. *Biol J Linn*
691 *Soc* 115, 522–531. doi:10.1111/bij.12532.
- 692 Johnston, A., Fink, D., Hochachka, W. M., and Kelling, S. (2018). Estimates of observer expertise
693 improve species distributions from citizen science data. *Methods in Ecology and Evolution* 9,
694 88–97.
- 695 Kelling, S., Johnston, A., Bonn, A., Fink, D., Ruiz-Gutierrez, V., Bonney, R., et al. (2019). Using
696 Semistructured Surveys to Improve Citizen Science Data for Monitoring Biodiversity.
697 *BioScience* 69, 170–179. doi:10.1093/biosci/biz010.
- 698 Kelling, S., Johnston, A., Hochachka, W. M., Iliff, M., Fink, D., Gerbracht, J., et al. (2015). Can
699 Observation Skills of Citizen Scientists Be Estimated Using Species Accumulation Curves?
700 *PLoS One* 10. doi:10.1371/journal.pone.0139600.
- 701 Kepler, C. B., and Scott, J. M. (1981). Reducing bird count variability by training observers. *Studies*
702 *in Avian Biology* 6.

- 703 Koenker, R., and Hallock, K. F. (2001). Quantile Regression. *Journal of Economic Perspectives* 15,
704 143–156. doi:10.1257/jep.15.4.143.
- 705 Lewandowski, E., and Specht, H. (2015). Influence of volunteer and project characteristics on data
706 quality of biological surveys. *Conservation Biology* 29, 713–723. doi:10.1111/cobi.12481.
- 707 Link, W. A., Sauer, J. R., and Niven, D. K. (2006). A hierarchical model for regional analysis of
708 population change using Christmas Bird Count data, with application to the American Black
709 Duck. *The Condor* 108, 13–24.
- 710 McKinley, D. C., Miller-Rushing, A. J., Ballard, H. L., Bonney, R., Brown, H., Cook-Patton, S. C., et
711 al. (2017). Citizen science can improve conservation science, natural resource management,
712 and environmental protection. *Biological Conservation* 208, 15–28.
713 doi:10.1016/j.biocon.2016.05.015.
- 714 Miller-Rushing, A., Primack, R., and Bonney, R. (2012). The history of public participation in
715 ecological research. *Frontiers in Ecology and the Environment* 10, 285–290.
716 doi:10.1890/110278.
- 717 R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for
718 Statistical Computing, Vienna, Austria.
- 719 Robbins, C. S., and Stallcup, R. W. (1981). Problems in separating species with similar habits and
720 vocalizations.
- 721 Robinson, W. D., and Curtis, J. R. (2020). Creating benchmark measurements of tropical forest bird
722 communities in large plots. *Condor* 122. doi:10.1093/condor/duaa015.
- 723 Robinson, W. D., Lees, A. C., and Blake, J. G. (2018). Surveying tropical birds is much harder than
724 you think: a primer of best practices. *Biotropica* 50, 846–849. doi:10.1111/btp.12608.
- 725 Schuetz, J. G., and Johnston, A. (2019). Characterizing the cultural niches of North American birds.
726 *Proc Natl Acad Sci USA* 116, 10868–10873. doi:10.1073/pnas.1820670116.
- 727 Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., and Kelling, S. (2009). eBird: A
728 citizen-based bird observation network in the biological sciences. *Biological Conservation*
729 142, 2282–2292. doi:10.1016/j.biocon.2009.05.006.
- 730 Walker, J., and Taylor, P. (2017). Using eBird data to model population change of migratory bird
731 species. *Avian Conservation and Ecology* 12. doi:10.5751/ACE-00960-120104.
- 732 West, M. J. (1999). Stereological methods for estimating the total number of neurons and synapses:
733 issues of precision and bias. *Trends in Neurosciences* 22, 51–61. doi:10.1016/S0166-
734 2236(98)01362-9.
- 735 Wood, C., Sullivan, B., Iliff, M., Fink, D., and Kelling, S. (2011). eBird: Engaging Birders in Science
736 and Conservation. *PLoS Biol* 9. doi:10.1371/journal.pbio.1001220.

737 Young, B. E., Dodge, N., Hunt, P. D., Ormes, M., Schlesinger, M. D., and Shaw, H. Y. (2019). Using
738 citizen science data to support conservation in environmental regulatory contexts. *Biological*
739 *Conservation* 237, 57–62. doi:10.1016/j.biocon.2019.06.016.

740 Yu, J., Wong, W.-K., and Hutchinson, R. A. (2010). Modeling Experts and Novices in Citizen
741 Science Data for Species Distribution Modeling. in *2010 IEEE International Conference on*
742 *Data Mining*, 1157–1162. doi:10.1109/ICDM.2010.103.

743

744

745

746

747

748 Tables.

749

750 Table 1. Twenty species were included in the study. Scientific names, sequence, and short-hand
 751 codes follow American Ornithological Society (<http://checklist.aou.org/taxa>). See text for definitions
 752 of dabbler versus diver and dispersed versus aggregated foragers. Plumage sexual dichromatism was
 753 scored based on the period of year in which the species is most numerous at the study site: weak or
 754 no dichromatism (0) and moderate to strong dichromatism (1).

755

English name	Scientific name	Code	Dabbler (0) or Diver (1)	Dispersed (0) or aggregated (1)	Plumage dichromatism
Wood Duck	<i>Aix sponsa</i>	wodu	0	0	1
Cinnamon Teal	<i>Spatula cyanoptera</i>	cite	0	0	0
Northern Shoveler	<i>Spatula clypeata</i>	nsho	0	1	1
Gadwall	<i>Mareca strepera</i>	gadw	0	0	1
American Wigeon	<i>Mareca americana</i>	amwi	0	1	1
Mallard	<i>Anas platyrhynchos</i>	mall	0	0	1
Northern Pintail	<i>Anas acuta</i>	nopi	0	0	1
Green-winged Teal	<i>Anas crecca</i>	gwte	0	1	1
Canvasback	<i>Aythya valisineria</i>	canv	1	0	1

Ring-necked Duck	<i>Aythya collaris</i>	rndu	1	1	1
Lesser Scaup	<i>Aythya affinis</i>	lesc	1	0	1
Surf Scoter	<i>Melanitta perspicillata</i>	susc	1	0	0
Bufflehead	<i>Bucephala albeola</i>	buff	1	0	1
Hooded Merganser	<i>Lophodytes cucullatus</i>	home	1	0	0
Ruddy Duck	<i>Oxyura jamaicensis</i>	rudu	1	1	0
Pied-billed Grebe	<i>Podilymbus podiceps</i>	pbgr	1	0	0
Eared Grebe	<i>Podiceps nigricollis</i>	eagr	1	0	0
American Coot	<i>Fulica americana</i>	amco	0	1	0
Ring-billed Gull	<i>Larus delawarensis</i>	rbgu	0	0	0
California Gull	<i>Larus californicus</i>	cagu	0	0	0

756

757

758

759

760

761

762

763

764

765 Table 2. Model results of variables most influential on percent error. R^2 is the quadratic of the daily
766 reference (benchmark) count; percent richness is the fraction of the waterbird species present each
767 day that were included on each eBird checklist; duration was the length (minutes) of eBird checklist
768 observation period; starttime was time of day each checklist was initiated; dichromatic was whether
769 each waterbird species exhibited plumage dichromatism or not; date2 was the quadratic of day of
770 year; and proRichness was the total species detected by WDR on each date. See supplemental
771 materials for the full model results.

772

	df	Log likelihood	BIC	delta	weight
R*2_Percent Richness	7	-9751.3	19565.0	0	0.696
R*2_Percent Richness_duration	8	-9749.0	19569.4	4.44	0.075
R*2_Percent Richness_starttime	8	-9749.2	19570.1	4.72	0.066
R*2_Percent Richness_dichromatic	8	-9749.4	19570.4	5.19	0.052
R*2_Percent Richness_date2	9	-9745.0	19572.7	5.41	0.047
R*2_Percent Richness_proRichness	8	-9750.7	19573.0	7.79	0.014

773

774

775

776 Supplemental Table 1. Species-specific measurements of central tendency and variation in percent
777 counting errors. A) excluding species non-detections from checklists; B) including species non-
778 detections (zero counts) in checklists.

779

780

781 Supplemental Table 2. Species-specific BIC model results. Full model results are presented for each
782 species alphabetically.

783

784 Supplemental Table 3. Full mixed-effects model results supplementing the abbreviated results
785 presented in Table 2.

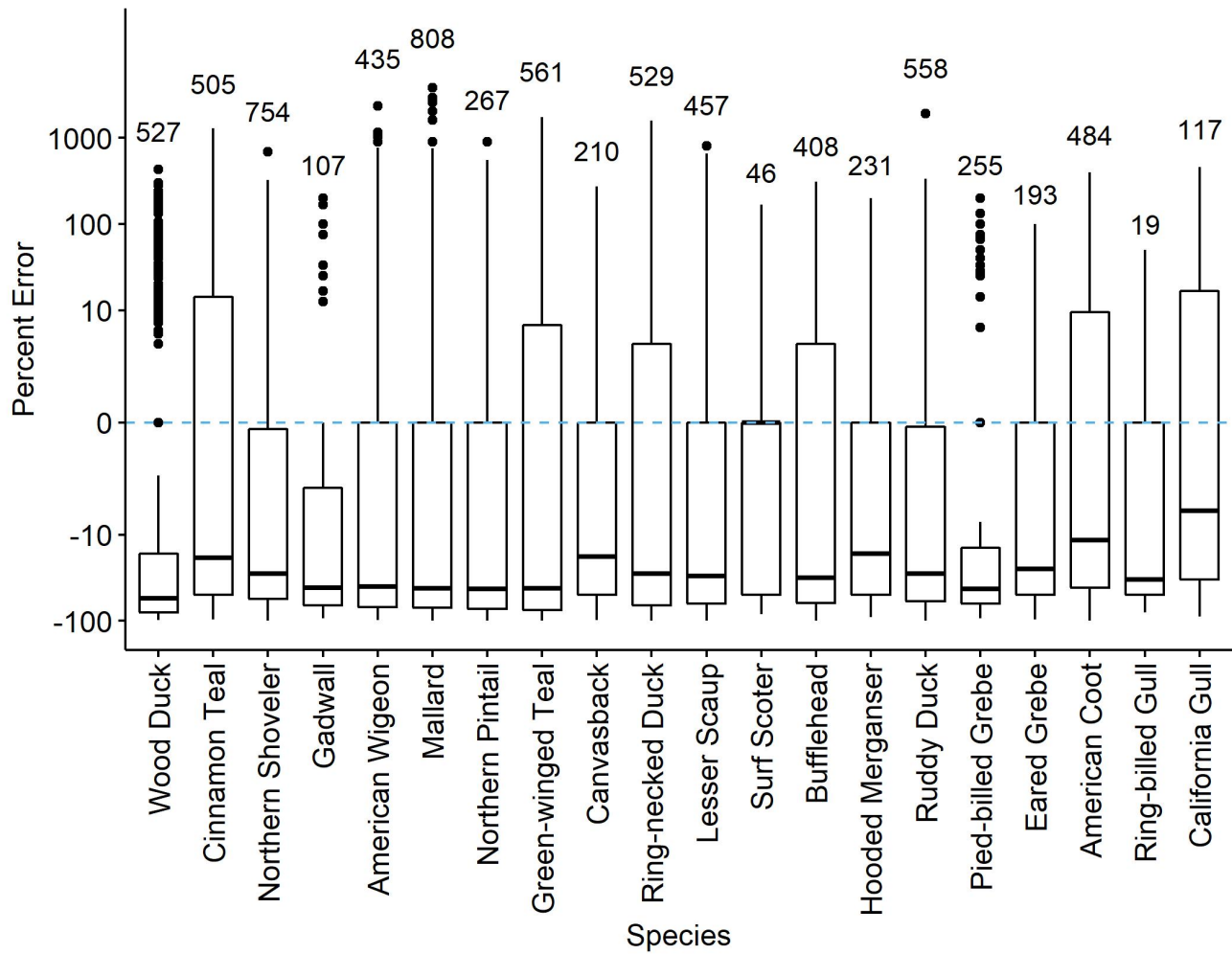
786

787

788 Figures.

789

790 A



791

792

793

794

795

796

797

798

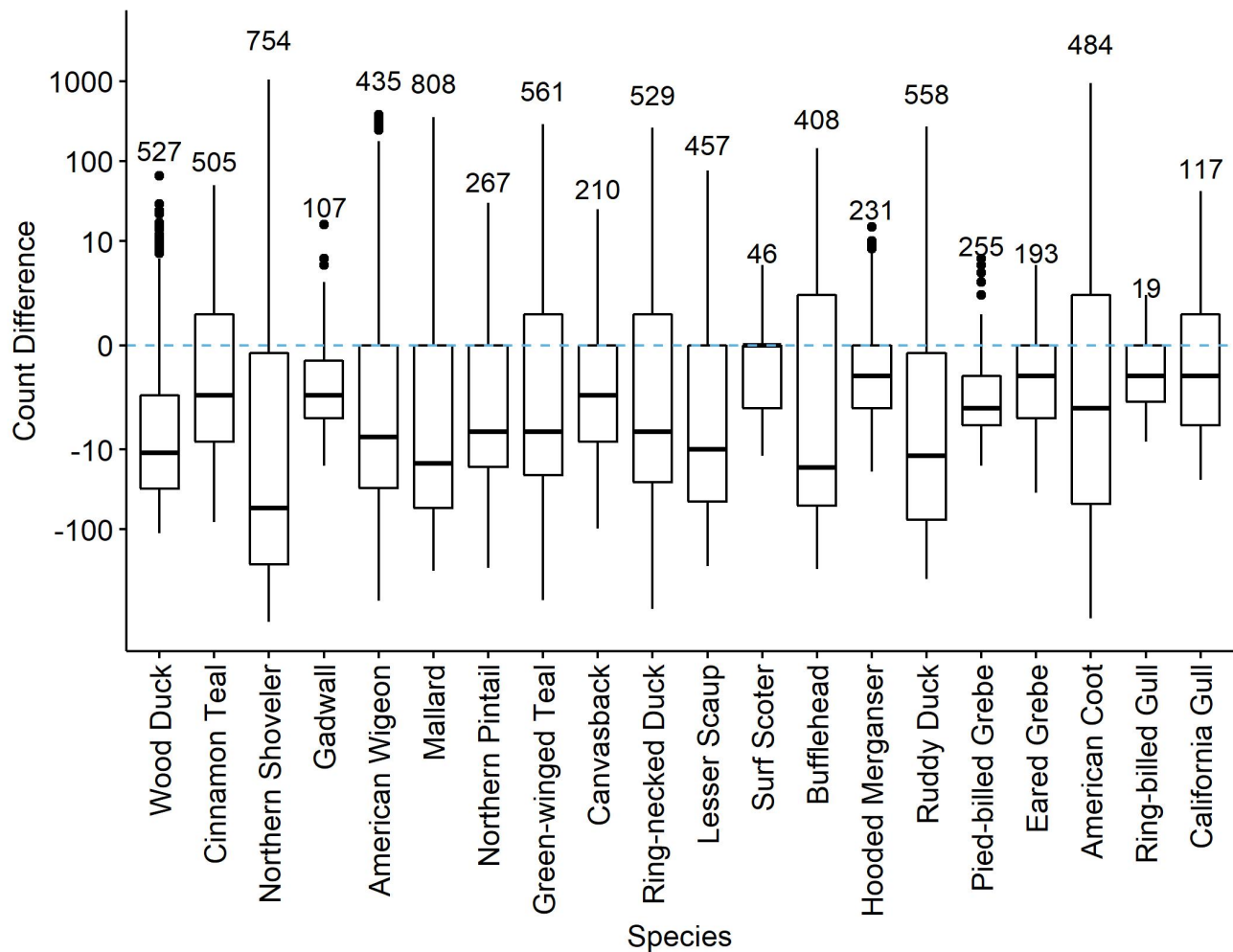
799

800 B

801

802

803



804

805 Figure 1. Percent error (A) and count differences (B) in counts of 20 waterbird species reported on
806 eBird checklists at the Philomath Ponds, Oregon USA, 2010-2019. Medians, quantile plots and
807 outliers are indicated, as well as number of checklists reporting counts of each species. Only
808 checklists reporting counts greater than zero were included. For checklists including counts of zero
809 on dates when R^* counts were non-zero, see Supplemental Figure 2.

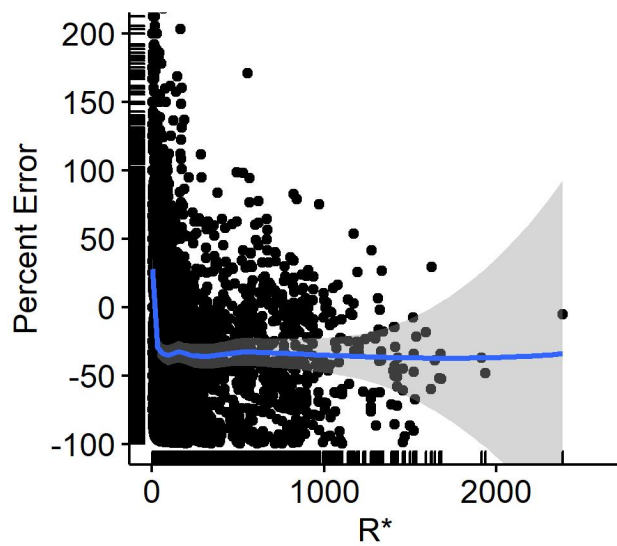
810

811

812

813 A

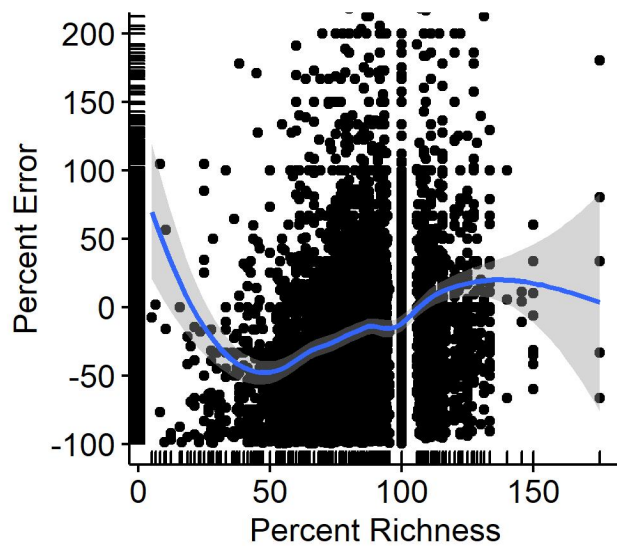
814



815

816 B

817



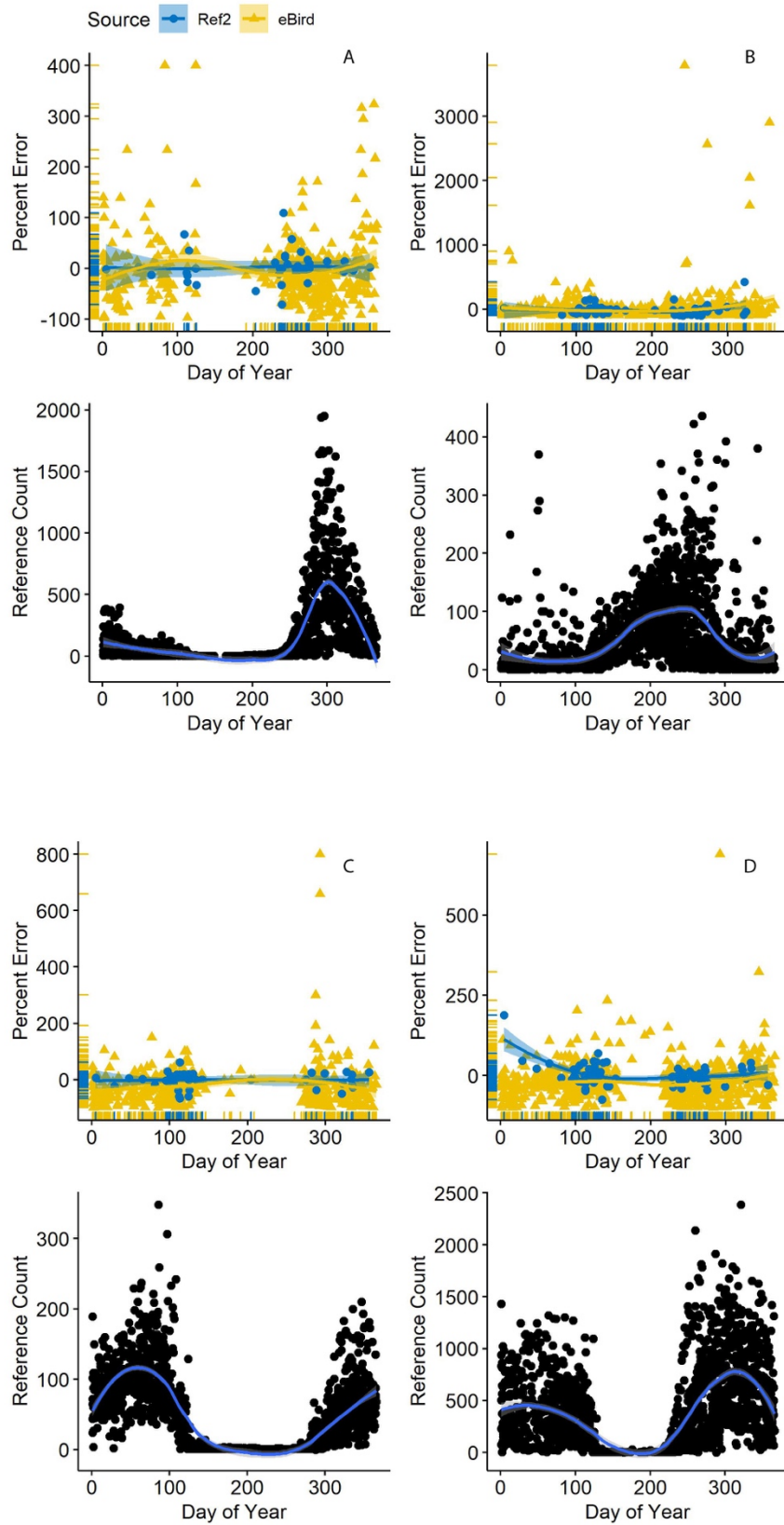
818

819 Figure 2. BIC-model predicted percent error in eBird waterbird counts as a function of A) reference
820 (benchmark) counts (R^*), and B) percent richness of waterbird species detected at Philomath ponds.

821

822

823

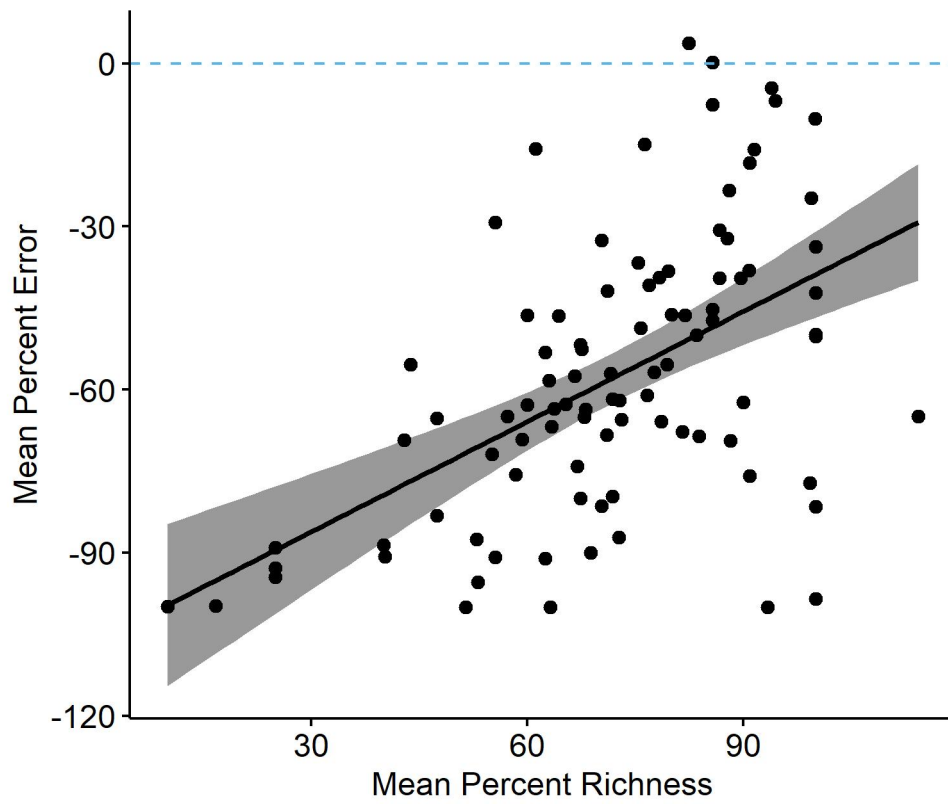


825 Figure 3. Variation in reference (benchmark) counts (R^*) as a function of date (lower panel) and
826 counts reported in eBird (gold triangles in upper panel) alongside second-visit counts (Ref2; blue
827 circles) at Philomath ponds, Oregon USA, 2010-2019. Counts in the upper panels are indicated with
828 respect to the R^* count (zero line) each day. Loess regression lines with 95% confidence intervals are
829 included. A) American Coot; B) Mallard; C) Lesser Scaup; D) Northern Shoveler.

830

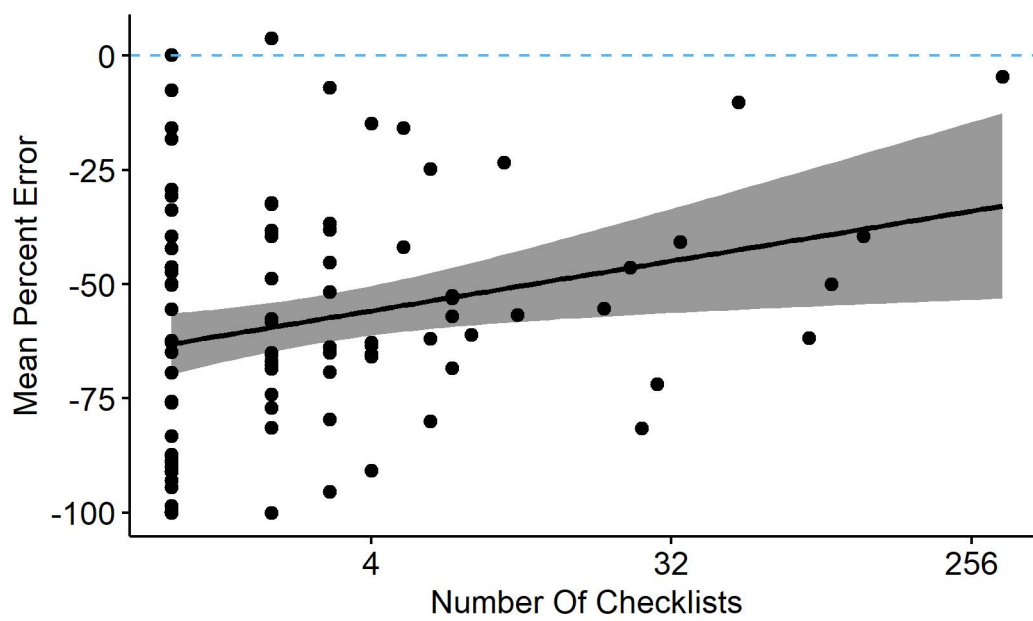
831

832 A



833

834 B



835

836

837

838 Figure 4. A) Observers reporting a greater percentage of waterbird species present at Philomath
839 ponds, Oregon USA, tended to have lower percent counting errors in their eBird checklists (linear
840 regression and 95% confidence intervals; $y = -110 + 0.68x$). B) Observers submitting more total
841 checklists tended to have lower counting errors ($y = -60 + 0.17x$). Note that these are means of all
842 applicable checklists for each observer, so each point represents a unique observer.

843

844

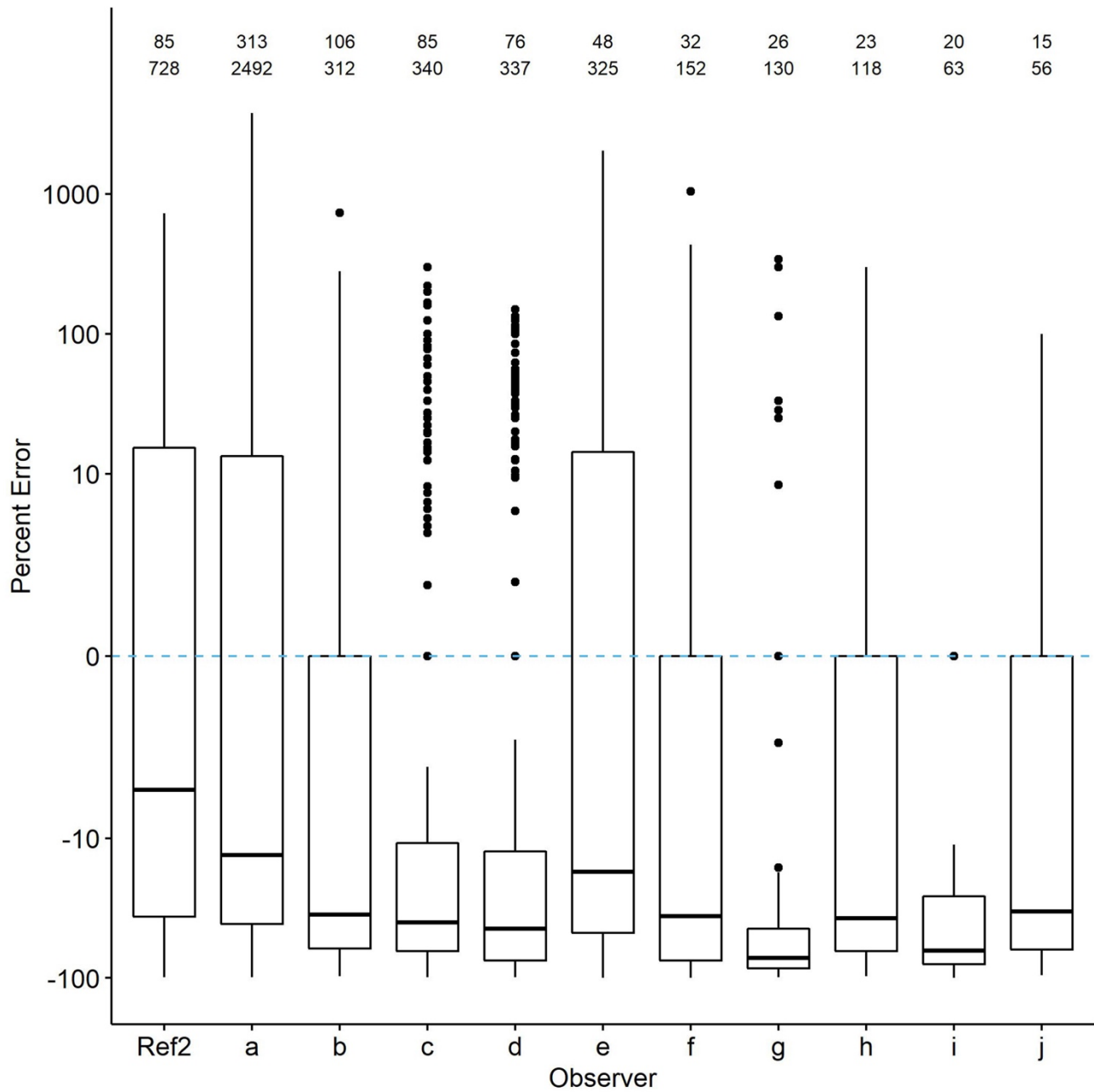
845

846

847

848

849



850

851 Figure 5. Comparison of percent count errors in eBird checklists contributed by the 10 observers with
852 the most checklists (top row of numbers) and waterbird observations (second row of numbers; each
853 checklist includes multiple species). The zero line is R^* . Ref2 is the second-visit data from WDR.
854 Quantile plots show the median, 25th percentiles as boxes and whiskers, plus outliers. Species-
855 specific plots are available from the authors upon request.

856

857

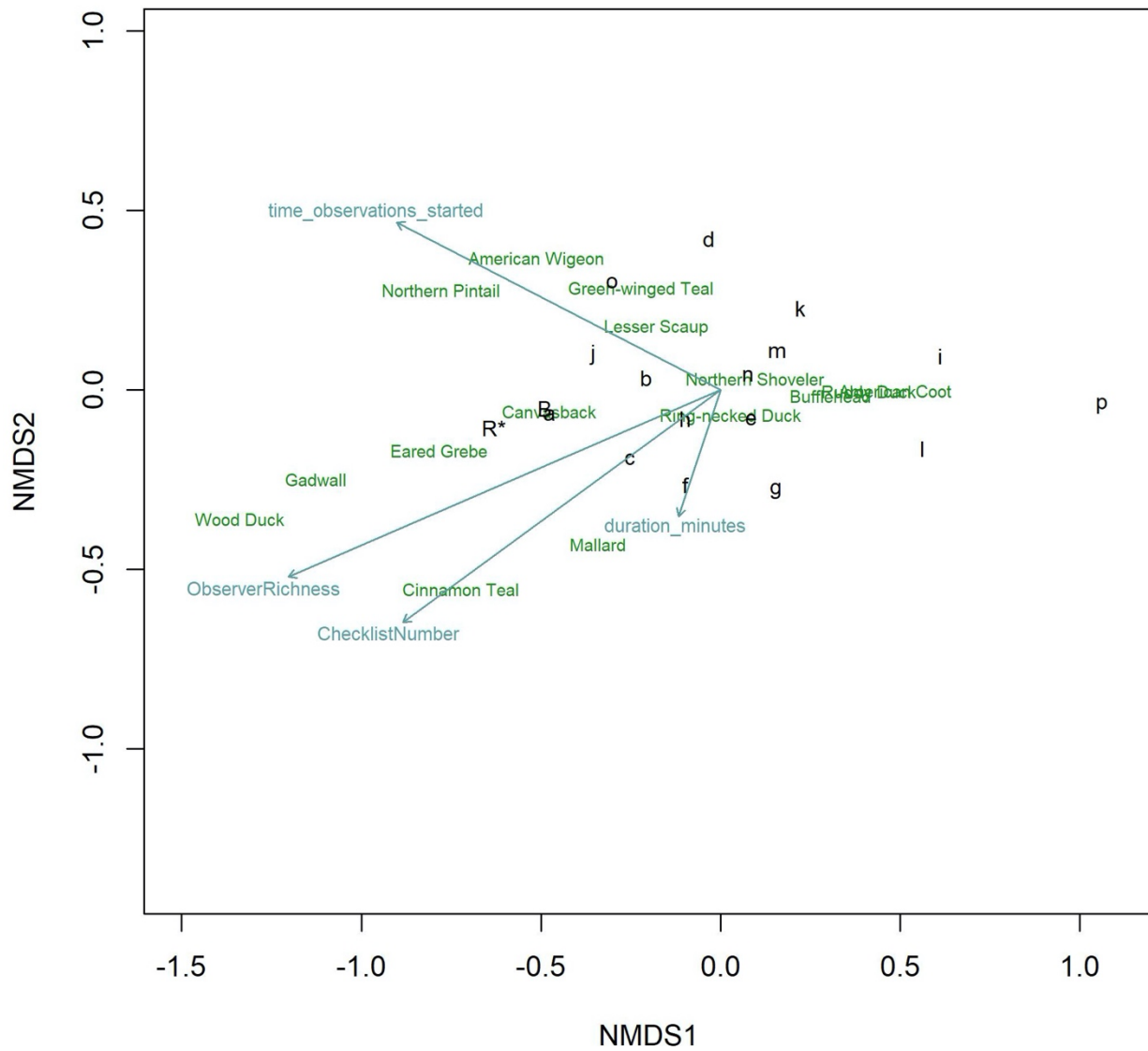
858

859

860 A

861

862



863

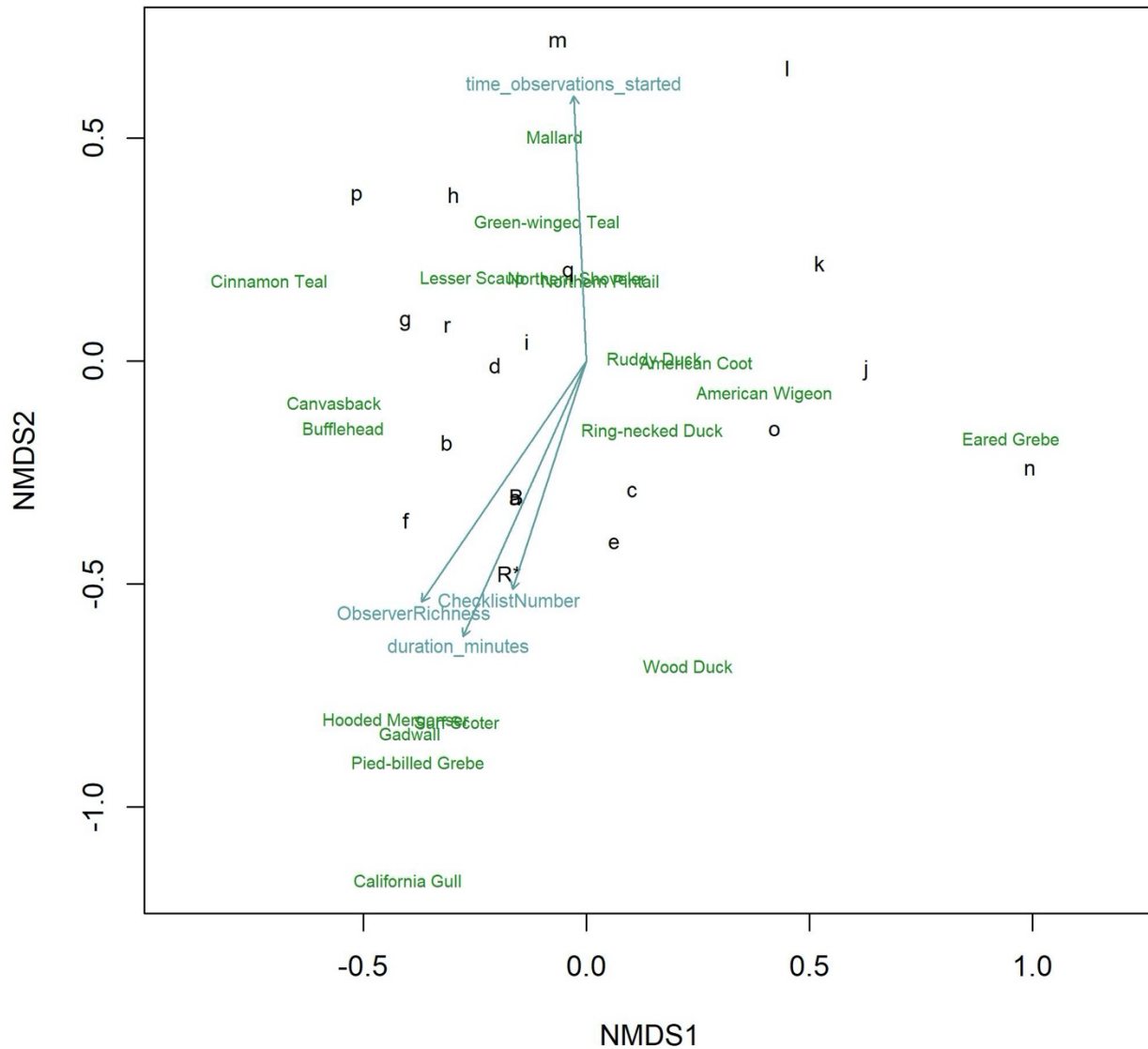
864

865

866

867

868 B

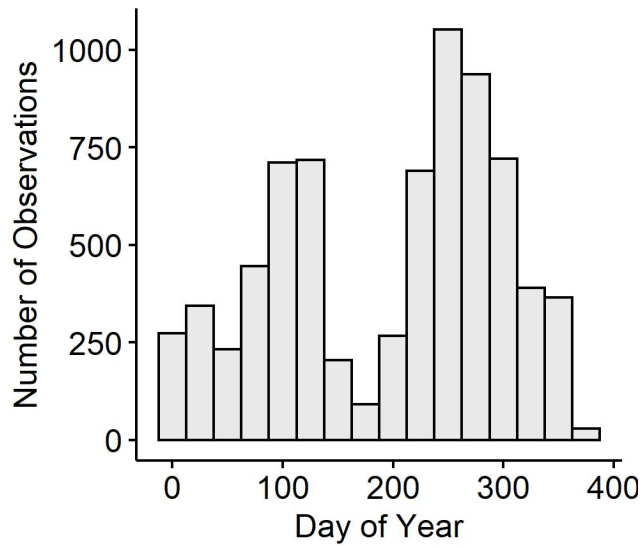


869
 870 Figure 6. Ordinations using NMDS of eBird checklists characterization of the waterbird community
 871 during A) January and B) October at Philomath ponds, Oregon USA, 2010-2019. The most
 872 influential vectors included Observer Richness (percent of known richness reported on each
 873 checklist), Checklist Number (total number of checklists per observer), observation start time each
 874 day, and the duration of each observation period. Relative positions of species in species space are
 875 noted by species English names. Benchmark counts are noted by R^* . Individual observers are noted
 876 by lower case letters; those nearest to R^* produced characterizations of the waterbird community
 877 most like R^* . \bar{B} is the collective average of eBird checklists, showing that from the perspective of
 878 generally characterizing the community, averaging across checklists contributed by many observers
 879 aligns more closely with R^* than do checklists from most individual observers, although observer *a*
 880 occupies nearly the same location in species space.

881 Supplemental Figures

882

883



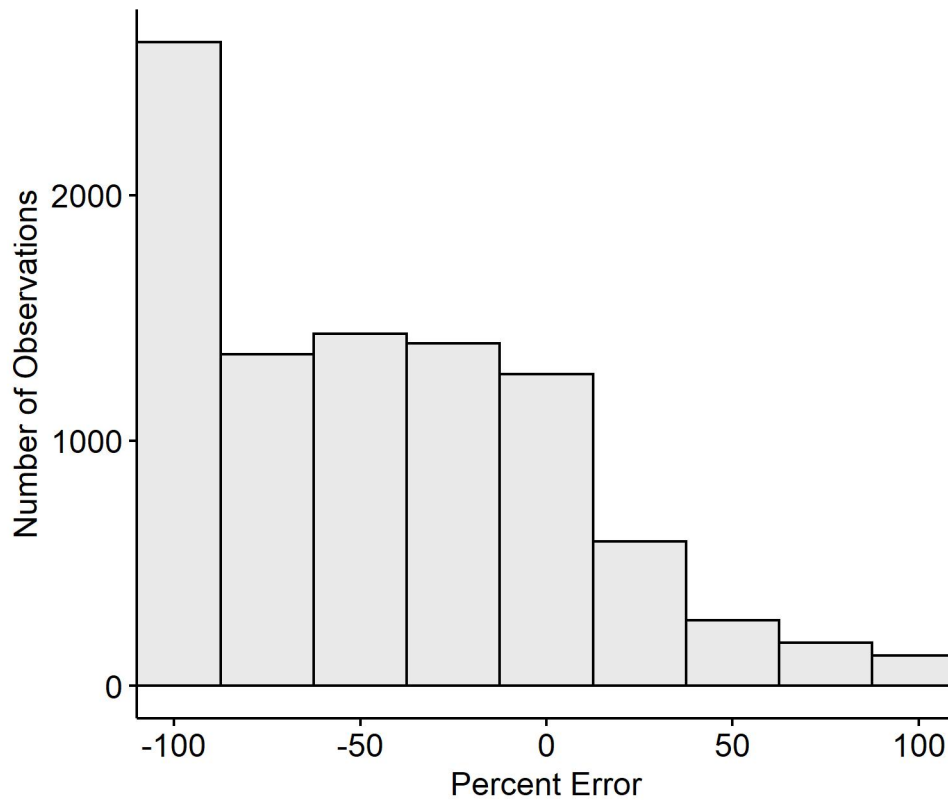
884

885 Supplemental Figure 1. Number of eBird checklists contributed for the study site at Philomath Ponds,
886 Oregon USA, 2010-2019, as a function of day of year.

887

888

889

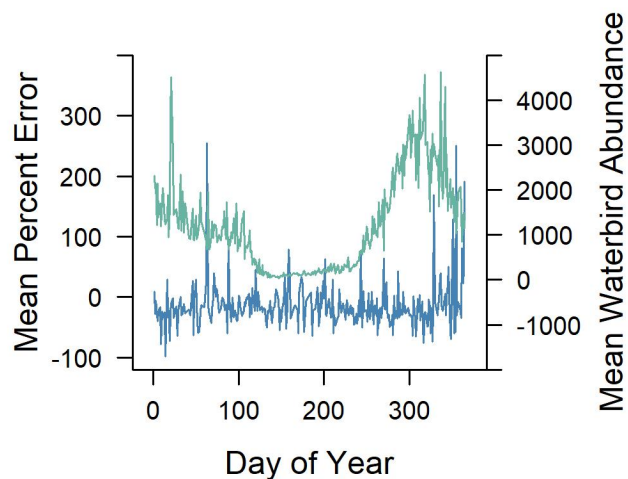


890

891 Supplemental Figure 2. Counts of waterbirds in eBird checklists included in our analyses as a
892 function of their percent error.

893

894



895

896 Supplemental Figure 3. Relationship between mean percent error on eBird checklists (blue line) and
897 mean waterbird abundance (green line) as a function of day of year at Philomath ponds, Oregon
898 USA, 2010-2019. Waterbird abundance is the mean of all the counts (R^*) of all of the possible 20
899 study species present each day across the 10 years.

900

901