

1   Forecasting biodiversity in breeding birds  
2                   using best practices

3                   *David J. Harris*<sup>1</sup> (*corresponding author*)

4                   *Shawn D. Taylor*<sup>2</sup>

5                   *Ethan P. White*<sup>1</sup>

6   <sup>1</sup> Department of Wildlife Ecology and Conservation, University of Florida, Gainesville,  
7   FL, United States

8   <sup>2</sup> School of Natural Resources and Environment, University of Florida Gainesville, FL,  
9   United States

## Abstract

Biodiversity forecasts are important for conservation, management, and evaluating how well current models characterize natural systems. While the number of forecasts for biodiversity is increasing, there is little information available on how well these forecasts work. Most biodiversity forecasts are not evaluated to determine how well they predict future diversity, fail to account for uncertainty, and do not use time-series data that captures the actual dynamics being studied. We addressed these limitations by using best practices to explore our ability to forecast the species richness of breeding birds in North America. We used hindcasting to evaluate six different modeling approaches for predicting richness. Hindcasts for each method were evaluated annually for a decade at 1,237 sites distributed throughout the continental United States. All models explained more than 50% of the variance in richness, but none of them consistently outperformed a baseline model that predicted constant richness at each site. The best practices implemented in this study directly influenced the forecasts and evaluations. Stacked species distribution models and “naive” forecasts produced poor estimates of uncertainty and accounting for this resulted in these models dropping in the relative performance compared to other models. Accounting for observer effects improved model performance overall, but also changed the rank ordering of models because it did not improve the accuracy of the “naive” model. Considering the forecast horizon revealed that the prediction accuracy decreased across all models as the time horizon of the forecast increased. To facilitate the rapid improvement of biodiversity forecasts, we emphasize the value of specific best practices in making forecasts and evaluating forecasting methods.

## 33 **Introduction**

34 Forecasting the future state of ecological systems is increasingly important for planning  
35 and management, and also for quantitatively evaluating how well ecological models  
36 capture the key processes governing natural systems (Clark et al. 2001, Dietze 2017,  
37 Houlahan et al. 2017). Forecasts regarding biodiversity are especially important, due to  
38 biodiversity's central role in conservation planning and its sensitivity to anthropogenic  
39 effects (Cardinale et al. 2012, Díaz et al. 2015, Tilman et al. 2017). High-profile studies  
40 forecasting large biodiversity declines over the coming decades have played a large role  
41 in shaping ecologists' priorities (as well as those of policymakers; e.g. IPCC 2014), but  
42 it is inherently difficult to evaluate such long-term predictions before the projected  
43 biodiversity declines have occurred.

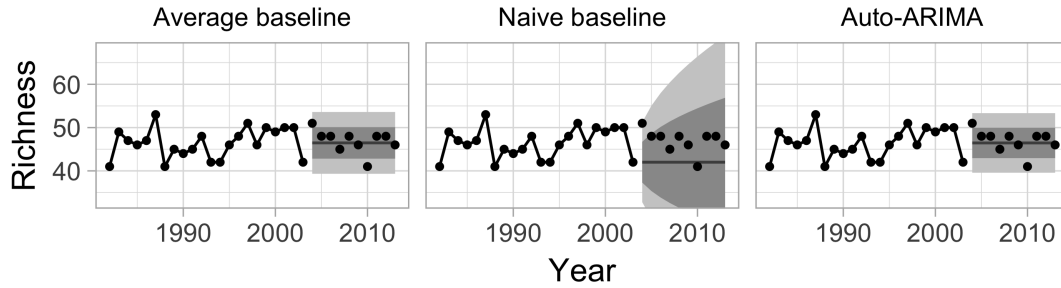
44 Previous efforts to predict future patterns of terrestrial species richness, and diversity  
45 more generally, have focused primarily on building species distributions models (SDMs;  
46 Thomas et al. 2004, Thuiller et al. 2011, Urban 2015). In general, these models  
47 describe individual species' occurrence patterns as functions of the environment. Given  
48 forecasts for environmental conditions, these models can predict where each species  
49 will occur in the future. These species-level predictions are then combined ("stacked")  
50 to generate forecasts for species richness (e.g. Calabrese et al. 2014). Alternatively,  
51 models that directly relate spatial patterns of species richness to environment conditions  
52 have been developed and generally perform equivalently to stacked SDMs (Algar et al.  
53 2009, Distler et al. 2015). This approach is sometimes referred to as "macroecological"  
54 modeling, because it models the larger-scale pattern (richness) directly (Distler et al.  
55 2015).

56 Despite the emerging interest in forecasting species richness and other aspects of  
57 biodiversity (Jetz et al. 2007, Thuiller et al. 2011), little is known about how effectively  
58 we can anticipate these dynamics. This is due in part to the long time scales over which  
59 many ecological forecasts are applied (and the resulting difficulty in assessing whether

60 the predicted changes occurred; Dietze et al. 2016). What we do know comes from a  
61 small number of hindcasting studies, where models are built from different time periods  
62 and evaluated on their ability to predict biodiversity patterns in contemporary (Algar et  
63 al. 2009, Distler et al. 2015) or historic (Blois et al. 2013, Maguire et al. 2016) periods  
64 not used for model fitting. These studies are a valuable first step, but lack several  
65 components that are important for developing forecasting models with high predictive  
66 accuracy, and for understanding how well different methods can predict the future.  
67 These “best practices” for effective forecasting and evaluation (Box 1) broadly involve:  
68 1) expanding the use of data to include biological and environmental time-series  
69 (Tredennick et al. 2016); 2) accounting for uncertainty in observations and processes,  
70 (Yu et al. 2010, Harris 2015); and 3) conducting meaningful evaluations of the forecasts  
71 by hindcasting, archiving short-term forecasts, and comparing forecasts to baselines to  
72 determine whether the forecasts are more accurate than assuming the system is basically  
73 static (Perretti et al. 2013).

74 In this paper, we attempt to forecast the species richness of breeding birds at over 1,200  
75 of sites located throughout North America, while following best practices for ecological  
76 forecasting (Box 1). To do this, we combine 32 years of time-series data on bird  
77 distributions from annual surveys with monthly time-series of climate data and  
78 satellite-based remote-sensing. Datasets that span a time scale of 30 years or more have  
79 only recently become available for large-scale time-series based forecasting. A dataset  
80 of this size allows us to model and assess changes a decade or more into the future in  
81 the presence of shifts in environmental conditions on par with predicted climate change.  
82 We compare traditional distribution modeling based approaches to spatial models of  
83 species richness, time-series methods, and two simple baselines that predict constant  
84 richness for each site, on average (Figure 1). All of our forecasting models account for  
85 uncertainty and observation error, are evaluated across different time lags using  
86 hindcasting, and are publicly archived to allow future assessment. We discuss the

## A. Single-site models



## B. Environmental models

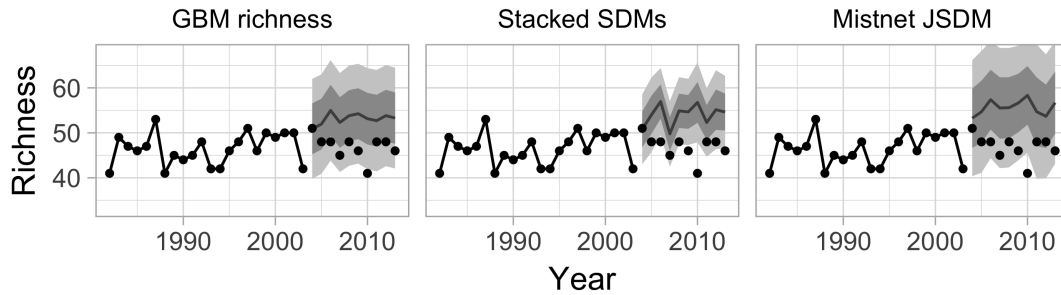


Figure 1: Example predictions from six forecasting models for a single site. Data from 1982 through 2003, connected by solid lines, were used for training the models; the remaining points were used for evaluating the models' forecasts. In each panel, point estimates for each year are shown with lines; the darker ribbon indicates the 68% prediction interval (1 standard deviation of uncertainty), and the lighter ribbon indicates the 95% prediction interval. **A.** Single-site models were trained independently on each site's observed richness values. The first two models ("average" and "naive") served as baselines. **B.** The environmental models were trained to predict richness based on elevation, climate, and NDVI; the environmental models' predictions change from year to year as environmental conditions change.

implications of these practices for our understanding of, and confidence in, the resulting forecasts, and how we can continue to build on these approaches to improve ecological forecasting in the future.

## Methods

We evaluated 6 types of forecasting models (Table 1) by dividing the 32 years of data into 22 years of training data and 10 years of data for evaluating forecasts using hindcasting. Here we use definitions from meteorology, where a hindcast is generally

any prediction for an event that has already happened, while forecasts are predictions for actual future events (Jolliffe and Stephenson 2003). We also made long term forecasts by using the full data set for training and making forecasts through the year 2050. For both time frames, we made forecasts using each model with and without correcting for observer effects, as described below.

## **Data**

**Richness data.** Bird species richness was obtained from the North American Breeding Bird Survey (BBS) (Pardieck et al. 2017) using the Data Retriever Python package (Morris and White 2013, Senyondo et al. 2017) and rdataretriever R package (McGlinn et al. 2017). BBS observations are three-minute point counts made at 50 fixed locations along a 40km route. Here we denote each route as a site and summarize richness as the total species observed at all 50 locations in each surveyed year. Prior to summarizing the data was filtered to exclude all nocturnal, cepuscular, and aquatic species (since these species are not well sampled by BBS methods; Hurlbert and White 2005), as well as unidentified species, and hybrids. All data from surveys that did not meet BBS quality criteria were also excluded.

We used observed richness values from 1982 (the first year of complete environmental data) to 2003 to train the models, and from 2004 to 2013 to test their performance. We only used BBS routes from the continental United States (i.e. routes where climate data was available PRISM Climate Group (2004)), and we restricted the analysis to routes that were sampled during 70% of the years in the training period (i.e., routes with at least 16 annual observations). The resulting dataset included 34,494 annual surveys of 1,279 unique sites, and included 385 species. Site-level richness varied from 8 to 91 with an average richness of 51 species.

**Past environmental data.** Environmental data included a combination of elevation, bioclimatic variables and a remotely sensed vegetation index (the normalized difference

120 vegetation index; NDVI), all of which are known to influence richness and distribution  
121 in the BBS data (Kent et al. 2014). For each year in the dataset, we used the 4 km  
122 resolution PRISM data (PRISM Climate Group 2004) to calculate eight bioclimatic  
123 variables identified as relevant to bird distributions (Harris 2015): mean diurnal range,  
124 isothermality, max temperature of the warmest month, mean temperature of the wettest  
125 quarter, mean temperature of the driest quarter, precipitation seasonality, precipitation  
126 of the wettest quarter, and precipitation of the warmest quarter. These variables were  
127 calculated for the 12 months leading up to the annual survey (July-June) as opposed to  
128 the calendar year. Satellite-derived NDVI, a primary correlate of richness in BBS data  
129 (Hurlbert and Haskell 2002), was obtained from the NDIV3g dataset with an 8 km  
130 resolution (Pinzon and Tucker 2014) and was available from 1981-2013. Average  
131 summer (April, May, June) and winter (December, January, February) NDVI values  
132 were used as predictors. Elevation was from the SRTM 90m elevation dataset (Jarvis et  
133 al. 2008) obtained using the R package raster (Hijmans 2016). Because BBS routes are  
134 40-km transects rather than point counts, we used the average value of each  
135 environmental variable within a 40 km radius of each BBS route's starting point.

136 **Future environmental projections.** In addition to the analyses presented here, we  
137 have also generated and archived long term forecasts from 2014-2050. This will allow  
138 future researchers to assess the performance of our six models on longer time horizons  
139 as more years of BBS data become available. Precipitation and temperature were  
140 forecast using the CMIP5 multi-model ensemble dataset (Brekke et al. 2013). 37  
141 downscaled model runs (Brekke et al. 2013, see Table S1) using the RCP6.0 scenario  
142 were averaged together to create a single ensemble used to calculate the bioclimatic  
143 variables for North America. For NDVI, we used the per-site average values from  
144 2000-2013 as a simple forecast. For observer effects (see below), each site was set to  
145 have zero observer bias. The predictions have been archived at (Harris et al. 2017b).

## 146 **Accounting for observer effects**

147 Observer effects are inherent in large data sets collected by different observers, and are  
148 known to occur in BBS (Sauer et al. 1994). For each forecasting approach, we trained  
149 two versions of the corresponding model: one with corrections for differences among  
150 observers, and one without (Figure 2). We estimated the observer effects (and  
151 associated uncertainty about those effects) using a linear mixed model, with observer as  
152 a random effect, built in the Stan probabilistic programming language (Carpenter et al.  
153 2017). Because observer and site are strongly related (observers tend to repeatedly  
154 sample the same site), site-level random effects were included to ensure that inferred  
155 deviations were actually observer-related (as opposed to being related to the sites that a  
156 given observer happened to see). The resulting model is described mathematically and  
157 with code in Supplement S1. The model partitions the variance in observed richness  
158 values into site-level variance, observer-level variance, and residual variance  
159 (e.g. variation within a site from year to year).

160 Across our six modeling approaches (described below), we used estimates from the  
161 observer model in three different ways. First, the expected values for site-level richness  
162 were used directly as our “average” baseline model (see below). For the two models that  
163 made species-level predictions, the estimated observer effects were included alongside  
164 the environmental variables as predictors. Finally, we trained the remaining models to  
165 predict observer-corrected richness values (i.e. observed richness minus the observer  
166 effect, or the number of species that would have been recorded by a “typical” observer).  
167 Since the site-level and observer-level random effects are not known precisely, we  
168 represented the range of possible values using 500 Monte Carlo samples from the  
169 posterior distribution over these effects. Each downstream model was then trained 500  
170 times using different possible values for the random effects.



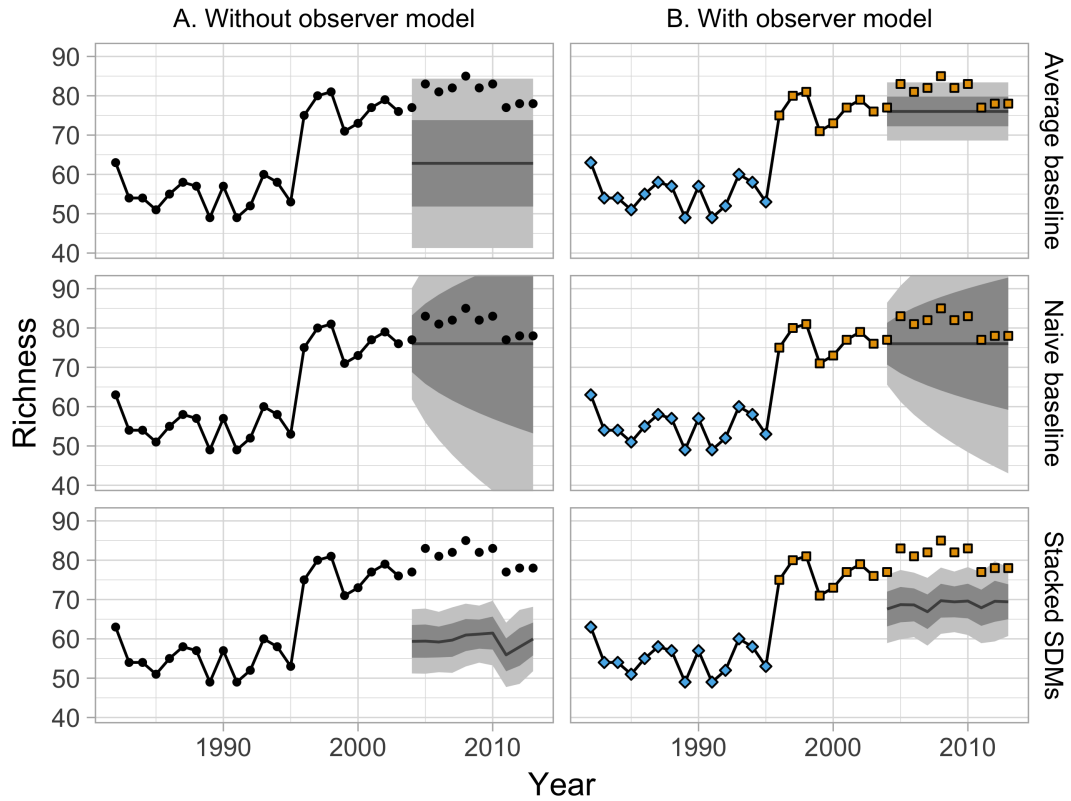


Figure 2: **A.** Model predictions for Pennsylvania route 35 when all observers are treated the same (black points). **B.** Model predictions for the same route when accounting for systematic differences between observers (represented by the points' colors). In this example most models are made more robust to observer turnover by including an observer model. Note that the “naive” model is less sensitive to observer turnover, and does not benefit as much from modeling it.

Table 1: Six forecasting models. Single-site models were trained site-by-site, without environmental data. Environmental models were trained at the continental scale, using only environmental variables (as opposed to site or time series information) as predictors. Most of the models were trained to predict richness directly. This mirrors the standard application of these techniques. Separate random forest SDMs were fit for each species and used to predict the probability of that species occurring at each site. The species-level probabilities at a site were summed to predict richness. The mistnet JSDM was trained to predict the full species composition at each site, and the number of species in its predictions was used as an estimate of richness.

Model	Response variable	Predictors		
		Site id	Time	Environment
Single-site models				
Average baseline	richness	✓		
Naive baseline	richness	✓	✓	
Auto-ARIMA	richness	✓	✓	
Environmental models				
GBM richness	richness			✓
Stacked SDMs	species-level presence			✓
Mistnet JSDM	species composition			✓

#### 171 **Models: site-level models**

172 Three of the models used in this study were fit to each site separately, with no  
 173 environmental information (Table 1). These models were fit to each BBS route twice:  
 174 once using the residuals from the observer model, and once using the raw richness  
 175 values. When correcting for observer effects, we averaged across 500 models that were  
 176 fit separately to the 500 Monte Carlo estimates of the observer effects, to account for  
 177 our uncertainty in the true values of those effects. All of these models use a Gaussian  
 178 error distribution (rather than a count distribution) for reasons discussed below (see  
 179 “Model evaluation”).

180 **Baseline models.** We used two simple baseline models as a basis for comparison with  
 181 the more complex models (Figure 2A). The first baseline, called the “average” model,  
 182 treated site-level richness observations as uncorrelated noise around a site-level

183 constant:

$$y_t = \mu + \epsilon_t.$$

184 Predictions from the “average” model are thus centered on  $\mu$ , which could either be the  
185 mean of the raw training richness values, or an output from the observer model. This  
186 model’s confidence intervals have a constant width that depends on the standard  
187 deviation of  $\epsilon$ , which can either be the standard deviation of the raw training richness  
188 values, or  $\sigma^{\text{residual}}$  from the observer model; see supplement).

189 The second baseline, called the “naive” model (Hyndman and Athanasopoulos 2014),  
190 was a simple autoregressive process with a single year of history, i.e. an ARIMA(0,1,0)  
191 model:

$$y_t = y_{t-1} + \epsilon_t,$$

192 where the standard deviation of  $\epsilon$  is a free parameter for each site. In contrast to the  
193 “average” model, whose predictions are based on the average richness across the whole  
194 time series, the “naive” model predicts that future observations will be similar to the  
195 *final* observed value (e.g., in our hindcasts the value observed in 2003). Moreover,  
196 because the  $\epsilon$  values accumulate over time, the confidence intervals expand rapidly as  
197 the predictions extend farther into the future. Despite these differences, both models’  
198 richness predictions are centered on a constant value, so neither model can anticipate  
199 any trends in richness or any responses to future environmental changes.

200 **Time series models.** We used Auto-ARIMA models (based on the `auto.arima`  
201 function in the package `forecast`; Hyndman 2017) to represent an array of different  
202 time-series modeling approaches. These models can include an autoregressive  
203 component (as in the “naive” model, but with the possibility of longer-term

dependencies in the underlying process), a moving average component (where the noise can have serial autocorrelation) and an integration/differencing component (so that the analysis could be performed on sequential differences of the raw data, accommodating more complex patterns including trends). The `auto.arima` function chooses whether to include each of these components (and how many terms to include for each one) using AICc (Hyndman 2017). Since there is no seasonal component to the BBS time-series, we did not include a season component in these models. Otherwise we used the default settings for this function (See supplement for details).

## **Models: environmental models**

In contrast to the single-site models, most attempts to predict species richness focus on using correlative models based on environmental variables. We tested three common variants of this approach: direct modeling of species richness; stacking individual species distribution models; and joint species distribution models (JSDMs). Following the standard approach, site-level random effects were not included in these models as predictors, meaning that this approach implicitly assumes that two sites with identical Bioclim, elevation, and NDVI values should have identical richness distributions. As above, we included observer effects and the associated uncertainty by running these models 500 times (once per MCMC sample).

**“Macroecological” model: richness GBM.** We used a boosted regression tree model using the `gbm` package (Ridgeway *et al.* 2017) to directly model species richness as a function of environmental variables. Boosted regression trees are a form of tree-based modeling that work by fitting thousands of small tree-structured models sequentially, with each tree optimized to reduce the error of its predecessors. They are flexible models that are considered well suited for prediction (Elith *et al.* 2008). This model was optimized using a Gaussian likelihood, with a maximum interaction depth of 5, shrinkage of 0.015, and up to 10,000 trees. The number of trees used for prediction was

230 selected using the “out of bag” estimator; this number averaged 6,700 for the  
231 non-observer data and 7,800 for the observer-corrected data.

232 **Species Distribution Model: stacked random forests.** Species distribution models  
233 (SDMs) predict individual species’ occurrence probabilities using environmental  
234 variables. Species-level models are used to predict richness by summing the predicted  
235 probability of occupancy across all species at a site. This avoids known problems with  
236 the use of thresholds for determining whether or not a species will be present at a site  
237 (Pellissier et al. 2013, Calabrese et al. 2014). Following Calabrese et al. (2014), we  
238 calculated the uncertainty in our richness estimate by treating richness as a sum over  
239 independent Bernoulli random variables:  $\sigma_{richness}^2 = \sum_i p_i(1 - p_i)$ , where  $i$  indexes  
240 species. By itself, this approach is known to underestimate the true community-level  
241 uncertainty because it ignores the uncertainty in the species-level probabilities  
242 (Calabrese et al. 2014). To mitigate this problem, we used an ensemble of 500 estimates  
243 for each of the species-level probabilities instead of just one, propagating the  
244 uncertainty forward. We obtained these estimates using random forests (Liaw and  
245 Wiener 2002), a common approach in the species distribution modeling literature.  
246 Random forests are constructed by fitting hundreds of independent regression trees to  
247 randomly-perturbed versions of the data (Cutler et al. 2007, Caruana et al. 2008). When  
248 correcting for observer effects, each of the 500 trees in our species-level random forests  
249 used a different Monte Carlo estimate of the observer effects as a predictor variable.

250 **Joint Species Distribution Model: mistnet.** Joint species distribution models  
251 (JSDMs) are a new approach that makes predictions about the full composition of a  
252 community instead of modeling each species independently as above (Warton et al.  
253 2015). JSDMs remove the assumed independence among species and explicitly account  
254 for the possibility that a site will be much more (or less) suitable for birds in general (or  
255 particular groups of birds) than one would expect based on the available environmental  
256 measurements alone. As a result, JSDMs do a better job of representing uncertainty

257 about richness than stacked SDMs (Harris 2015, Warton et al. 2015). We used the  
258 `mistnet` package (Harris 2015) because it is the only JSMD that describes species'  
259 environmental associations with nonlinear functions.

## 260 **Model evaluation**

261 We defined model performance for all models in terms of continuous Gaussian errors,  
262 instead of using discrete count distributions. Variance in species richness within sites  
263 was lower than predicted by several common count models, such as the Poisson or  
264 binomial (i.e. richness was underdispersed for individual sites), so these count models  
265 would have had difficulty fitting the data (cf. Calabrese et al. 2014). The use of a  
266 continuous distribution is adequate here, since richness had a relatively large mean (51)  
267 and all models produce continuous richness estimates. When a model was run multiple  
268 times for the purpose of correcting for observer effects, we used the mean of those runs'  
269 point estimates as our final point estimate and we calculated the uncertainty using the  
270 law of total variance (i.e.  $\text{Var}(\bar{y}) + \mathbb{E} [\text{Var}(y)]$ , or the variance in point estimates plus  
271 the average residual variance).

272 We evaluated each model's forecasts using the data for each year between 2004 and  
273 2013. We used three metrics for evaluating performance: 1) root-mean-square error  
274 (RMSE) to determine how far, on average, the models' predictions were from the  
275 observed value; 2) the 95% prediction interval coverage to determine how well the  
276 models predicted the range of possible outcomes; and 3) deviance (i.e. negative 2 times  
277 the Gaussian log-likelihood) as an integrative measure of fit that incorporates both  
278 accuracy and uncertainty. In addition to evaluating forecast performance in general, we  
279 evaluated how performance changed as the time horizon of forecasting increased by  
280 plotting performance metrics against year. Finally, we decomposed each model's  
281 squared error into two components: the squared error associated with site-level means  
282 and the squared error associated with annual fluctuations in richness within a site. This

decomposition describes the extent to which each model's error depends on consistent differences among sites versus changes in site-level richness from year to year.

All analyses were conducted using R (R Core Team 2017). Primary R packages used in the analysis included dplyr (Wickham et al. 2017), tidyr (Wickham 2017), gimms (Detsch 2016), sp (Pebesma and Bivand 2005, Bivand et al. 2013), raster (Hijmans 2016), prism (PRISM Climate Group 2004), rdataretriever (McGlinn et al. 2017), forecast (Hyndman and Khandakar 2008, Hyndman 2017), git2r (Widgren and others 2016), ggplot (Wickham 2009), mistnet (Harris 2015), viridis (Garnier 2017), rstan (Stan Development Team 2016), yaml (Stephens 2016), purrr (Henry and Wickham 2017), gbm (Ridgeway *et al.* 2017), randomForest (Liaw and Wiener 2002). Code to fully reproduce this analysis is available on GitHub (<https://github.com/weecology/bbs-forecasting>) and archived on Zenodo (Harris et al. 2017a).

## Results

The site-observer mixed model found that 70% of the variance in richness in the training set could be explained by differences among sites, and 21% could be explained by differences among observers. The remaining 9% represents residual variation, where a given observer might report a different number of species in different years. In the training set, the residuals had a standard deviation of about 3.6 species. After correcting for observer differences, there was little temporal autocorrelation in these residuals (i.e. the residuals in one year explain 1.3% of the variance in the residuals of the following year), suggesting that richness was approximately stationary between 1982 and 2003.

When comparing forecasts for richness across sites all methods performed well (Figure 3; all  $R^2 > 0.5$ ). However SDMs (both stacked and joint) and the macroecological

308 model all failed to successfully forecast the highest-richness sites, resulting in a notable  
309 clustering of predicted values near ~60 species and the poorest model performance  
310 ( $R^2=0.52-0.78$ , versus  $R^2=0.67-0.87$  for the within-site methods).

311 While all models generally performed well in absolute terms (Figure 3), none  
312 consistently outperformed the “average” baseline (Figure 4). The auto-ARIMA was  
313 generally the best-performing non-baseline model, but in many cases (67% of the time),  
314 the auto.arima procedure selected a model with only an intercept term (i.e. no  
315 autoregressive terms, no drift, and no moving average terms), making it similar to the  
316 “average” model. All five alternatives to the “average” model achieved lower error on  
317 some of the sites in some years, but each one had a higher mean absolute error and  
318 higher mean deviance (Figure 4).

319 Most models produced confidence intervals that were too narrow, indicating  
320 overconfident predictions (Figure 5C). The random forest-based SDM stack was the  
321 most overconfident model, with only 72% of observations falling inside its 95%  
322 confidence intervals. This stacked SDM’s narrow predictive distribution caused it to  
323 have notably higher deviance (Figure 5B) than the next-worst model, even though its  
324 point estimates were not unusually bad in terms of RMSE (5A). As discussed elsewhere  
325 (Harris 2015), this overconfidence is a product of the assumption in stacked SDMs that  
326 errors in the species-level predictions are independent. The GBM-based  
327 “macroecological” model and the mistnet JSMD had the best calibrated uncertainty  
328 estimates (Figure 5B) and therefore their relative performance was higher in terms of  
329 deviance than in terms of RMSE. The “naive” model was the only model whose  
330 confidence intervals were too wide (Figure 5C), which can be attributed to the rapid rate  
331 at which these intervals expand (Figure 1).

332 Partitioning each model’s squared error shows that the majority of the residual error was  
333 attributed to errors in estimating site-level means, rather than errors in tracking  
334 year-to-year fluctuations (Figure 6). The “average” model, which was based entirely on



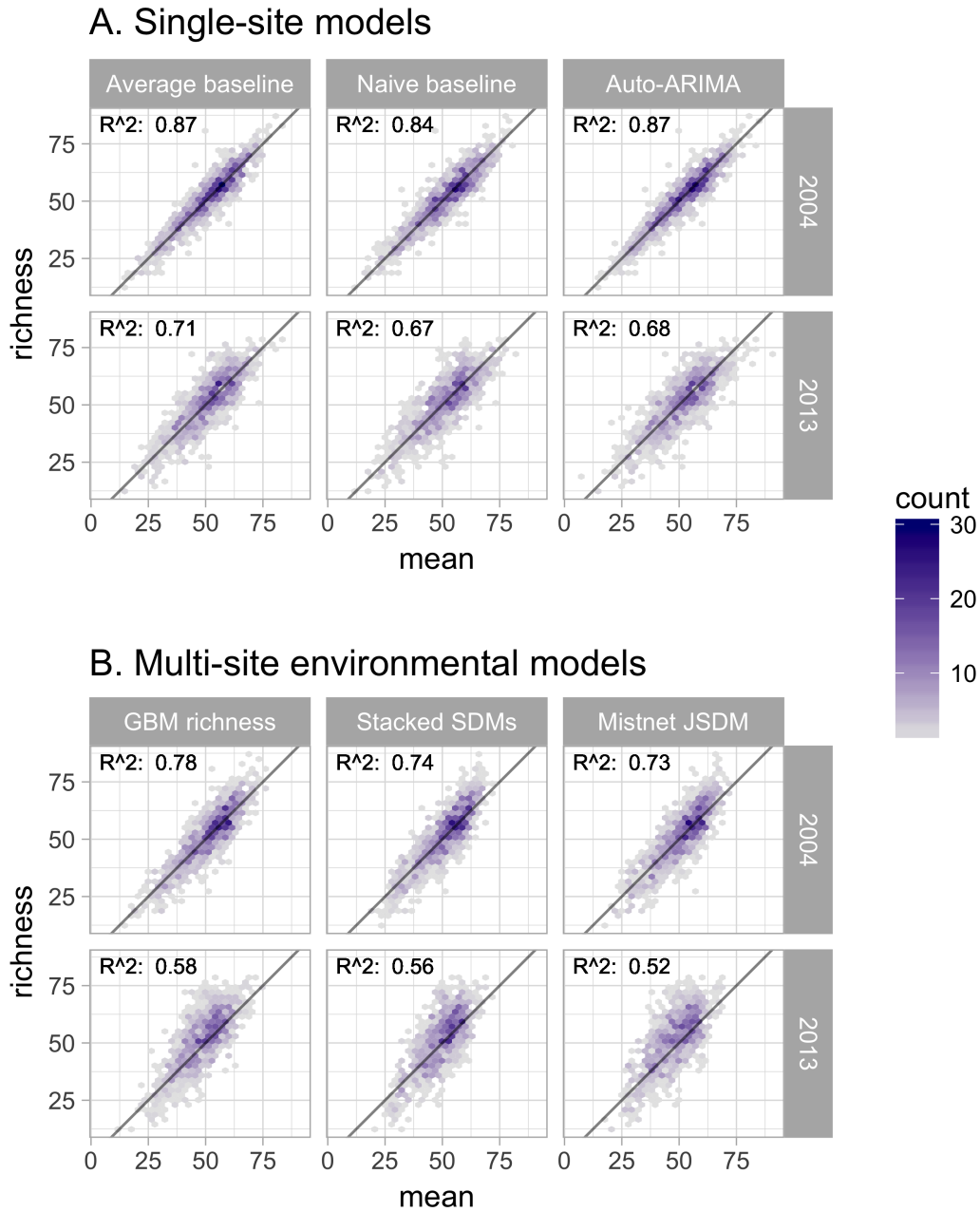


Figure 3: Performance of six forecasting models for predicting species richness one year (2004) and ten years into the future (2013). Plots show observed vs. predicted values for species richness. Models were trained with data from 1982-2003. In general, the single-site models (A) outperformed the environmental models (B). The accuracy of the predictions generally declined as the timescale of the forecast was extended from 2004 to 2013.

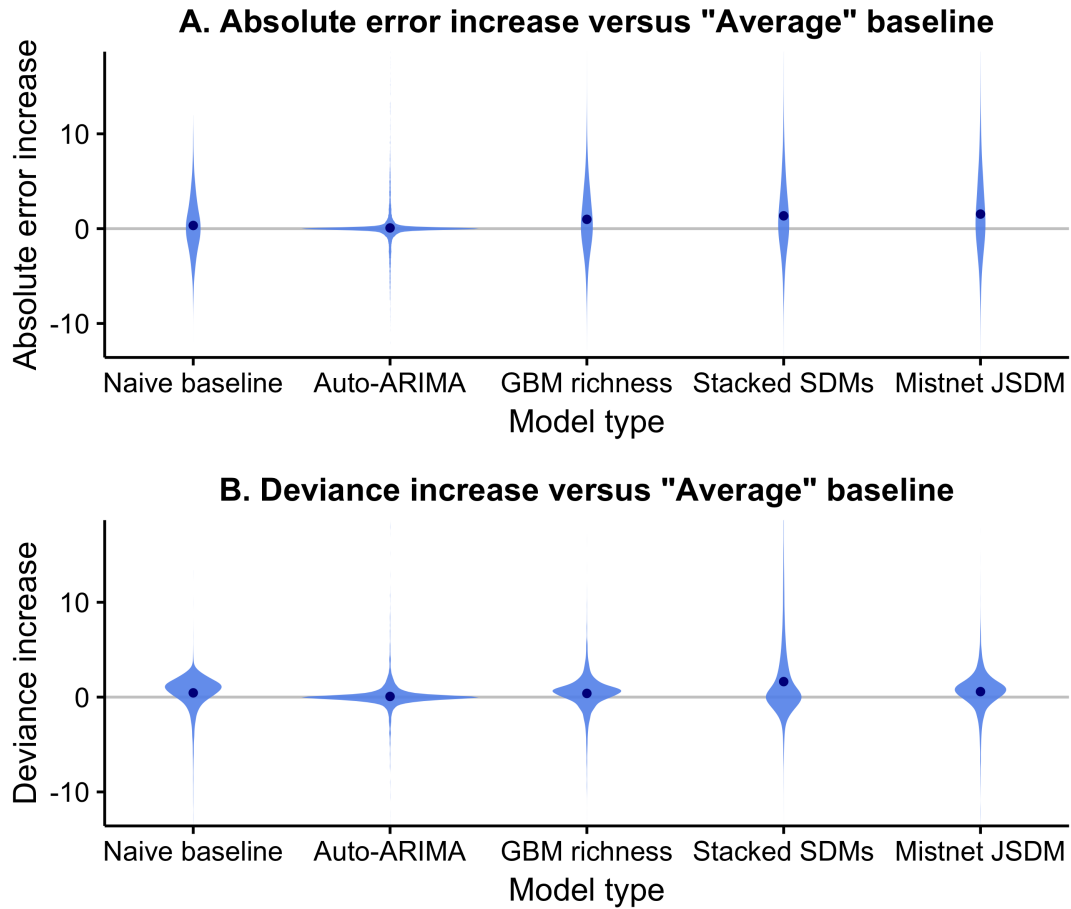


Figure 4: Difference between the forecast error of models and the error of the average baseline using both absolute error (A.) and deviance (B.). Differences are taken for each site and testing year so that errors for the same forecast are directly compared. The error of the average baseline is by definition zero and is indicated by the horizontal gray line. None of the five models provided a consistent improvement over the average baseline. The absolute error of the models was generally similar or larger than that of the “average” model, with large outliers in both directions. The deviance of the models was also generally higher than the “average” baseline.

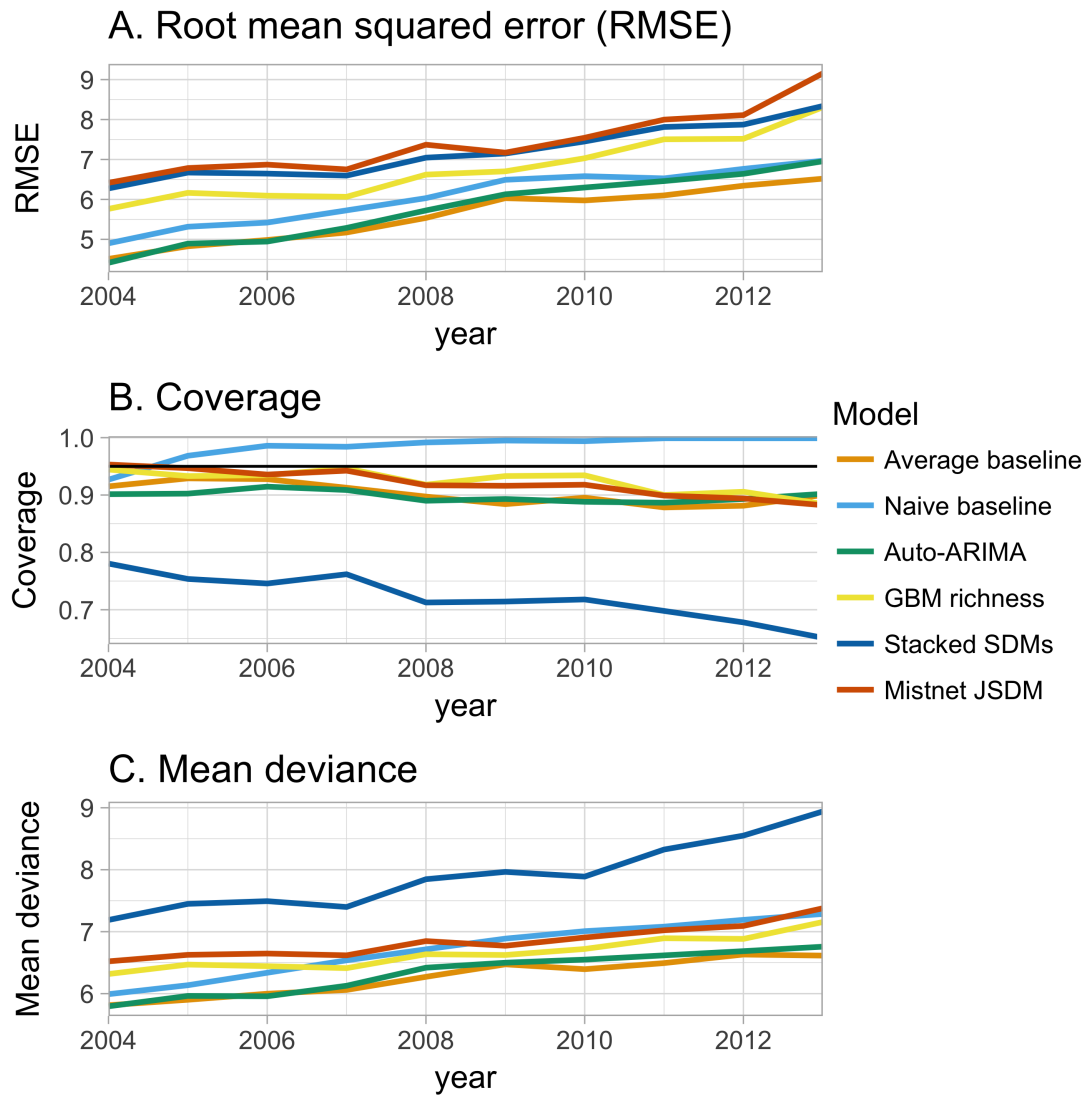


Figure 5: Change in performance of the six forecasting models with the time scale of the forecast (1-10 years into the future). **A.** Root mean square error (rmse; the error in the point estimates) shows the three environmental models tending to show the largest errors at all time scales and the models getting worse as they forecast further into the future at approximately the same rate. **B.** Deviance (lack of fit of the entire predictive distribution) shows the stacked species distribution models with much higher error than other models and shows that the “naive” model’s deviance grows relatively quickly. **C.** Coverage of a model’s 95% confidence intervals (how often the observed values fall inside the predicted range; the black line indicates ideal performance) shows that the “naive” model’s predictive distribution is too wide (capturing almost all of the data) and the stacked SDM’s predictive distribution is too narrow (missing almost a third of the observed richness values by 2014).

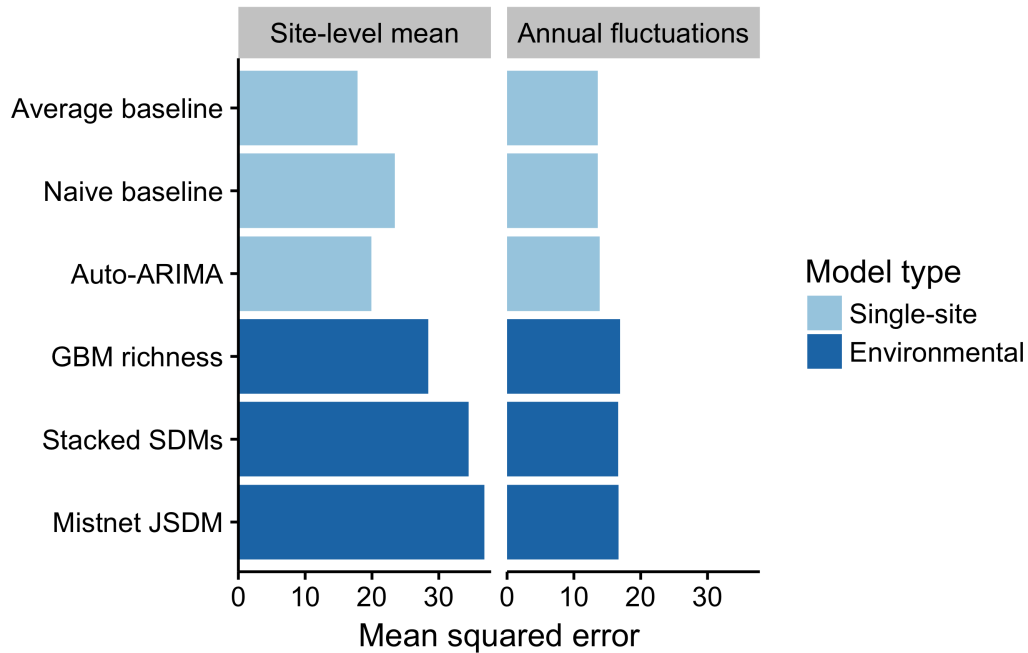


Figure 6: Partitioning of the squared error for each model into site and year components. The site-level mean component shows consistent over or under estimates of richness at a site across years. The annual fluctuation component shows errors in predicting fluctuations in a site’s richness over time. Both components of the mean squared error were lower for the single-site models than for the environmental models.

335 site-level means, had the lowest error in this regard. In contrast, the three environmental  
 336 models showed larger biases at the site level, though they still explained most of the  
 337 variance in this component. This makes sense, given that they could not explicitly  
 338 distinguish among sites with similar climate, NDVI, and elevation. Interestingly, the  
 339 environmental models had higher squared error than the baselines did for tracking  
 340 year-to-year fluctuations in richness as well.

341 Accounting for differences among observers generally improved measures of model fit  
 342 (Figure 7). Improvements primarily resulted from a small number of forecasts where  
 343 observer turnover caused a large shift in the reported richness values. The naive  
 344 baseline was less sensitive to these shifts, because it largely ignored the richness values  
 345 reported by observers that had retired by the end of the training period (Figure 1). The  
 346 average model, which gave equal weight to observations from the whole training period,

347 showed a larger decline in performance when not accounting for observer effects –  
348 especially in terms of coverage. The performance of the mistnet JSDM was notable  
349 here, because its prediction intervals retained good coverage even when not correcting  
350 for observer differences, which we attribute to the JSDM’s ability to model this  
351 variation with its latent variables.

## 352 **Discussion**

353 Forecasting is an emerging imperative in ecology; as such, the field needs to develop  
354 and follow best practices for conducting and evaluating ecological forecasts (Clark et al.  
355 2001). We have used a number of these practices (Box 1) in a single study that builds  
356 and evaluates forecasts of biodiversity in the form of species richness. The results of  
357 this effort are both promising and humbling. When comparing predictions across sites,  
358 many different approaches produce reasonable forecasts (Figure 3). If a site is predicted  
359 to have a high number of species in the future, relative to other sites, it generally does.  
360 However, none of the methods evaluated reliably determined how site-level richness  
361 changes over time (Figure 6), which is generally the stated purpose of these forecasts.  
362 As a result, baseline models, which did not attempt to anticipate changes in richness  
363 over time, generally provided the best forecasts for future biodiversity. While this study  
364 is restricted to breeding birds in North America, its results are consistent with a growing  
365 literature on the limits of ecological forecasting, as discussed below.

366 The most commonly used methods for forecasting future biodiversity, SDMs and  
367 macroecological models, both produced worse forecasts than time-series models and  
368 simple baselines. This weakness suggests that predictions about future biodiversity  
369 change should be viewed with skepticism unless the underlying models have been  
370 validated temporally, via hindcasting and comparison with simple baselines. Since  
371 site-level richness is relatively stable, spatial validation is not enough: a model can have  
372 high accuracy across spatial gradients without being able to predict changes over time.

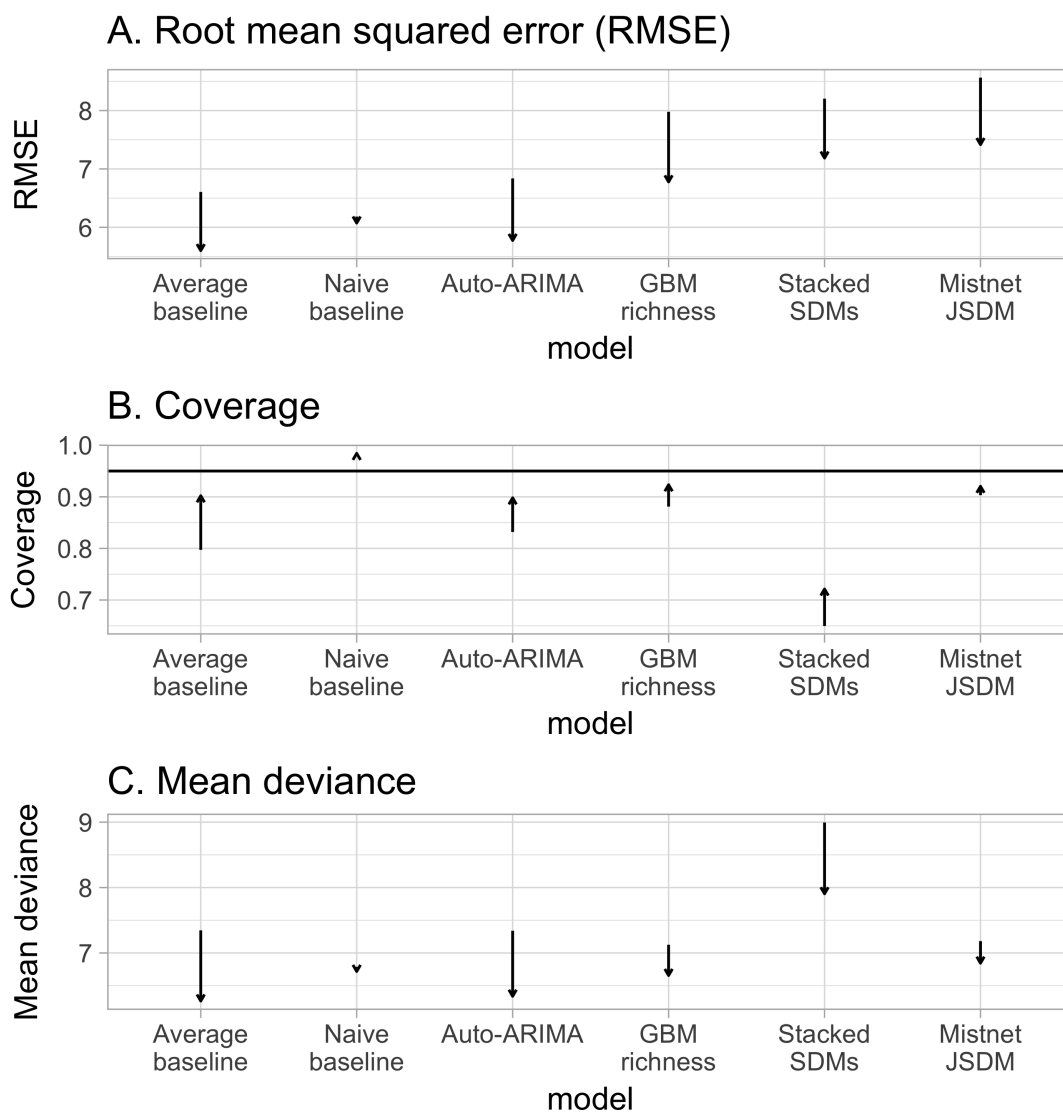


Figure 7: Controlling for differences among observers generally improved each model's predictions, on average. The magnitude of this effect was negligible for the Naive baseline, however.

373 This gap between spatial and temporal accuracy is known to be important for  
374 species-level predictions (Rapacciuolo et al. 2012, Oedekoven et al. 2017); our results  
375 indicate that it is substantial for higher-level patterns like richness as well. SDMs' poor  
376 temporal predictions are particularly sobering, as these models have been one of the  
377 main foundations for estimates of the predicted loss of biodiversity to climate change  
378 over the past decade or so (Thomas et al. 2004, Thuiller et al. 2011, Urban 2015). Our  
379 results also highlight the importance of comparing multiple modeling approaches when  
380 conducting ecological forecasts, and in particular, the value of comparing results to  
381 simple baselines to avoid over-interpreting the information present in these forecasts  
382 [Box 1]. Disciplines that have more mature forecasting cultures often do this by  
383 reporting "forecast skill", i.e., the improvement in the forecast relative to a simple  
384 baseline (Jolliffe and Stephenson 2003). We recommend following the example of  
385 Perretti et al. (2013) and adopting this approach in future ecological forecasting  
386 research.

387 When comparing different methods for forecasting our results demonstrate the  
388 importance of considering uncertainty (Box 1; Clark et al. 2001, Dietze et al. 2016).  
389 Previous comparisons between stacked SDMs and macroecological models reported  
390 that the methods yielded equivalent results for forecasting diversity (Algar et al. 2009,  
391 Distler et al. 2015). While our results support this equivalence for point estimates, they  
392 also show that stacked SDMs dramatically underestimate the range of possible  
393 outcomes; after ten years, more than a third of the observed richness values fell outside  
394 the stacked SDMs' 95% prediction intervals. Consistent with Harris (2015) and Warton  
395 et al. (2015), we found that JSDMs' wider prediction intervals enabled them to avoid  
396 this problem. Macroecological models appear to share this advantage, while being  
397 considerably easier to implement.

398 We have only evaluated annual forecasts up to a decade into the future, but forecasts are  
399 often made with a lead time of 50 years or more. These long-term forecasts are difficult

400 to evaluate given the small number of century-scale datasets, but are important for  
401 understanding changes in biodiversity at some of the lead times relevant for  
402 conservation and management. Two studies have assessed models of species richness at  
403 longer lead times (Algar et al. 2009, Distler et al. 2015), but the results were not  
404 compared to baseline or time-series models (in part due to data limitations) making  
405 them difficult to compare to our results directly. Studies on shorter time scales, such as  
406 ours, provide one way to evaluate our forecasting methods without having to wait  
407 several decades to observe the effects of environmental change on biodiversity (Petchey  
408 et al. 2015, Dietze et al. 2016, Tredennick et al. 2016), but cannot fully replace  
409 longer-term evaluations (Tredennick et al. 2016). In general, drivers of species richness  
410 can differ at different temporal scales (Rosenzweig 1995, White 2004, 2007, Blonder et  
411 al. 2017), so different methods may perform better for different lead times. In particular,  
412 we might expect environmental and ecological information to become more important  
413 at longer time scales, and thus for the performance of simple baseline forecasts to  
414 degrade faster than forecasts from SDMs and other similar models. We did observe a  
415 small trend in this direction: deviance for the auto-ARIMA models and for the average  
416 baseline grew faster than for two of the environmental models (the JSMD and the  
417 macroecological model), although this growth was not statistically significant for the  
418 average baseline.

419 While it is possible that models that include species' relationships to their environments  
420 or direct environmental constraints on richness will provide better fits at longer lead  
421 times, it is also possible that they will continue to produce forecasts that are worse than  
422 baselines that assume the systems are static. This would be expected to occur if richness  
423 in these systems is not changing over the relevant multi-decadal time scales, which  
424 would make simpler models with no directional change more appropriate. Recent  
425 suggestions that local scale richness in some systems is not changing directionally at  
426 multi-decadal scales supports this possibility (Brown et al. 2001, Ernest and Brown



2001, Vellend et al. 2013, Dornelas et al. 2014). A lack of change in richness may be expected even in the presence of substantial changes in environmental conditions and species composition at a site due to replacement of species from the regional pool (Brown et al. 2001, Ernest and Brown 2001). On average, the Breeding Bird Survey sites used in this study show little change in richness (site-level SD of 3.6 species, after controlling for differences among observers; see also La Sorte and Boecklen 2005). The absence of rapid change in this dataset is beneficial for the absolute accuracy of forecasts across different sites: when a past year's richness is already known, it is easy to estimate future richness. Ward et al. (2014) found similar patterns in time series of fisheries stocks, where relatively stable time series were best predicted by simple models and more complex models were only beneficial with dynamic time series. The site-level stability of the BBS data also explains why SDMs and macroecological models perform relatively well at predicting future richness, despite failing to capture changes in richness over time.

The relatively stable nature of the BBS richness time-series also makes it difficult to improve forecasts relative to simple baselines, since those baselines are already close to representing what is actually occurring in the system. It is possible that in systems exhibiting directional changes in richness and other biodiversity measures that models based on spatial patterns may yield better forecasts. Future research in this area should determine if regions or time periods exhibiting strong directional changes in biodiversity are better predicted by these models and also extend our forecast horizon analyses to longer timescales where possible. Our results also suggest that future efforts to understand and forecast biodiversity should incorporate species composition, since lower-level processes are expected to be more dynamic (Ernest and Brown 2001, Dornelas et al. 2014) and contain more information about how the systems are changing (Harris 2015). More generally, determining the forecastability of different aspects of ecological systems under different conditions is an important next step for the future of

454 ecological forecasting.

455 Future biodiversity forecasting efforts also need to address the uncertainty introduced  
456 by the error in forecasting the environmental conditions that are used as predictor  
457 variables. In this, and other hindcasting studies, the environmental conditions for the  
458 “future” are known because the data has already been observed. However, in real  
459 forecasts the environmental conditions themselves have to be predicted, and  
460 environmental forecasts will also have uncertainty and bias. Ultimately, ecological  
461 forecasts that use environmental data will therefore be more uncertain than our current  
462 hindcasting efforts, and it is important to correctly incorporate this uncertainty into our  
463 models (Clark et al. 2001, Dietze 2017). Limitations in forecasting future  
464 environmental conditions—particularly at small scales—will present continued  
465 challenges for models incorporating environmental variables, and this may result in a  
466 continued advantage for simple single-site approaches.

467 In addition to comparing and improving the process models used for forecasting it is  
468 important to consider the observation models. When working with any ecological  
469 dataset, there are imperfections in the sampling process that have the potential to  
470 influence results. With large scale surveys and citizen science datasets, such as the  
471 Breeding Bird Survey, these issues are potentially magnified by the large number of  
472 different observers and by major differences in the habitats and species being surveyed  
473 (Sauer et al. 1994). Accounting for differences in observers reduced the average error in  
474 our point estimates and also improved the coverage of the confidence intervals. In  
475 addition, controlling for observer effects resulted in changes in which models performed  
476 best, most notably improving most models’ point estimates relative to the naive baseline.  
477 This demonstrates that modeling observation error can be important for properly  
478 estimating and reducing uncertainty in forecasts and can also lead to changes in the best  
479 methods for forecasting [Box 1]. This suggests that, prior to accounting for observer  
480 effects, the naive model performed well largely because it was capable of

481 accommodating rapid shifts in estimated richness introduced by changes in the observer.  
482 These kinds of rapid changes were difficult for the other single-site models to  
483 accommodate. Another key aspect of an ideal observation model is imperfect detection.  
484 In this study, we did not address differences in detection probability across species and  
485 sites (Boulinier et al. 1998) since there is no clear way to address this issue using North  
486 American Breeding Bird Survey data without making strong assumptions about the data  
487 (i.e., assuming there is no biological variation in stops along a route; White and Hurlbert  
488 2010), but this would be a valuable addition to future forecasting models.

489 The science of forecasting biodiversity remains in its infancy and it is important to  
490 consider weaknesses in current forecasting methods in that context. In the beginning,  
491 weather forecasts were also worse than simple baselines, but these forecasts have  
492 continually improved throughout the history of the field (McGill 2012, Silver 2012,  
493 Bauer et al. 2015). One practice that lead to improvements in weather forecasts was that  
494 large numbers of forecasts were made publicly, allowing different approaches to be  
495 regularly assessed and refined (McGill 2012, Silver 2012). To facilitate this kind of  
496 improvement, it is important for ecologists to start regularly making and evaluating real  
497 ecological forecasts, even if they perform poorly, and to make these forecasts openly  
498 available for assessment (McGill 2012, Dietze et al. 2016). These forecasts should  
499 include both short-term predictions, which can be assessed quickly, and mid- to  
500 long-term forecasts, which can help ecologists to assess long time-scale processes and  
501 determine how far into the future we can successfully forecast (Dietze et al. 2016,  
502 Tredennick et al. 2016). We have openly archived forecasts from all six models through  
503 the year 2050 (Harris et al. 2017b), so that we and others can assess how well they  
504 perform. We plan to evaluate these forecasts and report the results as each new year of  
505 BBS data becomes available, and make iterative improvements to the forecasting  
506 models in response to these assessments.

507 Making successful ecological forecasts will be challenging. Ecological systems are

508 complex, our fundamental theory is less refined than for simpler physical and chemical  
509 systems, and we currently lack the scale of data that often produces effective forecasts  
510 through machine learning. Despite this, we believe that progress can be made if we  
511 develop an active forecasting culture in ecology that builds and assesses forecasts in  
512 ways that will allow us to improve the effectiveness of ecological forecasts more rapidly  
513 (Box 1; McGill 2012, Dietze et al. 2016). This includes expanding the scope of the  
514 ecological and environmental data we work with, paying attention to uncertainty in both  
515 model building and forecast evaluation, and rigorously assessing forecasts using a  
516 combination of hindcasting, archived forecasts, and comparisons to simple baselines.

## 517 **Acknowledgments**

518 This research was supported by the Gordon and Betty Moore Foundation's Data-Driven  
519 Discovery Initiative through Grant GBMF4563 to E.P. White. We thank the developers  
520 and providers of the data and software that made this research possible including: the  
521 PRISM Climate Group at Oregon State University, the staff at USGS and volunteer  
522 citizen scientists associated with the North American Breeding Bird Survey, NASA, the  
523 World Climate Research Programme's Working Group on Coupled Modelling and its  
524 working groups, the U.S. Department of Energy's Program for Climate Model  
525 Diagnosis and Intercomparison, and the Global Organization for Earth System Science  
526 Portals. A. C. Perry provided valuable comments that improved the clarity of this  
527 manuscript.

## 528 **Box 1: Best practices for making and evaluating ecological forecasts**

### 529 **1. Compare multiple modeling approaches**

530 Typically ecological forecasts use one modeling approach or a small number of related  
531 approaches. By fitting and evaluating multiple modeling approaches we can learn more  
532 rapidly about the best approaches for making predictions for a given ecological quantity  
533 (Clark et al. 2001, Ward et al. 2014). This includes comparing process-based (e.g.,  
534 Kearney and Porter 2009) and data-driven models (e.g., Ward et al. 2014), as well as  
535 comparing the accuracy of forecasts to simple baselines to determine if the modeled  
536 forecasts are more accurate than the naive assumption that the world is static (Jolliffe  
537 and Stephenson 2003, Perretti et al. 2013).

### 538 **2. Use time-series data when possible**

539 Forecasts describe how systems are expected to change through time. While some areas  
540 of ecological forecasting focus primarily on time-series data (Ward et al. 2014), others  
541 primarily focus on using spatial models and space-for-time substitutions (Blois et al.  
542 2013). Using ecological and environmental time-series data allows the consideration of  
543 actual dynamics from both a process and error structure perspective (Tredennick et al.  
544 2016).

### 545 **3. Pay attention to uncertainty**

546 Understanding uncertainty in a forecast is just as important as understanding the  
547 average or expected outcome. Failing to account for uncertainty can result in  
548 overconfidence in uncertain outcomes leading to poor decision making and erosion of  
549 confidence in ecological forecasts (Clark et al. 2001). Models should explicitly include  
550 sources of uncertainty and propagate them through the forecast where possible (Clark et

551 al. 2001, Dietze 2017). Evaluations of forecasts should assess the accuracy of models'  
552 estimated uncertainties as well as their point estimates (Dietze 2017).

#### 553 **4. Use predictors related to the question**

554 Many ecological forecasts use data that is readily available and easy to work with.  
555 While ease of use is a reasonable consideration it is also important to include predictor  
556 variables that are expected to relate to the ecological quantity being forecast.  
557 Time-series of predictors, instead of long-term averages, are also preferable to match  
558 the ecological data (see #2). Investing time in identifying and acquiring better predictor  
559 variables may have at least as many benefits as using more sophisticated modeling  
560 techniques (Kent et al. 2014).

#### 561 **5. Address unknown or unmeasured predictors**

562 Ecological systems are complex and many biotic and abiotic aspects of the environment  
563 are not regularly measured. As a result, some sites may deviate in consistent ways from  
564 model predictions. Unknown or unmeasured predictors can be incorporated in models  
565 using site-level random effects (potentially spatially autocorrelated) or by using latent  
566 variables that can identify unmeasured gradients (Harris 2015).

#### 567 **6. Assess how forecast accuracy changes with time-lag**

568 In general, the accuracy of forecasts decreases with the length of time into the future  
569 being forecast (Petchey et al. 2015). This decay in accuracy should be considered when  
570 evaluating forecasts. In addition to simple decreases in forecast accuracy the potential  
571 for different rates of decay to result in different relative model performance at different  
572 lead times should be considered.

## 573 **7. Include an observation model**

574 Ecological observations are influenced by both the underlying biological processes  
575 (e.g. resource limitation) and how the system is sampled. When possible, forecasts  
576 should model the factors influencing the observation of the data (Yu et al. 2010,  
577 Hutchinson et al. 2011, Schurr et al. 2012).

## 578 **8. Validate using hindcasting**

579 Evaluating a model's predictive performance across time is critical for understanding if  
580 it is useful for forecasting the future. Hindcasting uses a temporal out-of-sample  
581 validation approach to mimic how well a model would have performed had it been run  
582 in the past. For example, using occurrence data from the early 20th century to model  
583 distributions which are validated with late 20th century occurrences. Dense time series,  
584 such as yearly observations, are desirable to also evaluate the forecast horizon (see #6),  
585 but this is not a strict requirement.

## 586 **9. Publicly archive forecasts**

587 Forecast values and/or models should be archived so that they can be assessed after new  
588 data is generated (McGill 2012, Silver 2012, Dietze et al. 2016). Enough information  
589 should be provided in the archive to allow unambiguous assessment of each forecast's  
590 performance (Tetlock and Gardner 2016).

## 591 **10. Make both short-term and long-term predictions**

592 Even in cases where long-term predictions are the primary goal, short-term predictions  
593 should also be made to accommodate the time-scales of planning and management

594 decisions and to allow the accuracy of the forecasts to be quickly evaluated (Dietze et al.  
595 2016, Tredennick et al. 2016).

## 596 **References**

- 597 Algar, A. C., H. M. Kharouba, E. R. Young, and J. T. Kerr. 2009. Predicting the future  
598 of species diversity: Macroecological theory, climate change, and direct tests of  
599 alternative forecasting methods. *Ecography* 32:22–33.
- 600 Bauer, P., A. Thorpe, and G. Brunet. 2015. The quiet revolution of numerical weather  
601 prediction. *Nature* 525:47–55.
- 602 Bivand, R. S., E. Pebesma, and V. Gomez-Rubio. 2013. Applied spatial data analysis  
603 with R, second edition. Springer, NY.
- 604 Blois, J. L., J. W. Williams, M. C. Fitzpatrick, S. T. Jackson, and S. Ferrier. 2013. Space  
605 can substitute for time in predicting climate-change effects on biodiversity  
606 110:9374–9379.
- 607 Blonder, B., D. E. Moulton, J. Blois, B. J. Enquist, B. J. Graae, M. Macias-Fauria, B.  
608 McGill, S. Nogué, A. Ordonez, B. Sandel, and J.-C. Svenning. 2017. Predictability in  
609 community dynamics. *Ecology Letters* 20:293–306.
- 610 Boulinier, T., J. D. Nichols, J. R. Sauer, J. E. Hines, and K. Pollock. 1998. Estimating  
611 species richness: The importance of heterogeneity in species detectability. *Ecology*  
612 79:1018–1028.
- 613 Brekke, L., B. Thrasher, E. Maurer, and T. Pruitt. 2013. Downscaled cmip3 and cmip5  
614 climate and hydrology projections: Release of downscaled cmip5 climate projections,  
615 comparison with preceding information, and summary of user needs. US Dept. of the  
616 Interior, Bureau of Reclamation, Technical Services Center, Denver.
- 617 Brown, J. H., S. Ernest, J. M. Parody, and J. P. Haskell. 2001. Regulation of diversity:



618 Maintenance of species richness in changing environments. *Oecologia* 126.

619 Calabrese, J. M., G. Certain, C. Kraan, and C. F. Dormann. 2014. Stacking species  
620 distribution models and adjusting bias by linking them to macroecological models.  
621 *Global Ecology and Biogeography* 23:99–112.

622 Cardinale, B. J., J. E. Duffy, A. Gonzalez, D. U. Hooper, C. Perrings, P. Venail, A.  
623 Narwani, G. M. Mace, D. Tilman, D. A. Wardle, and others. 2012. Biodiversity loss and  
624 its impact on humanity. *Nature* 486:59–67.

625 Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M.  
626 Brubaker, J. Guo, P. Li, and A. Riddell. 2017. Stan : A Probabilistic Programming  
627 Language. *Journal of Statistical Software* 76.

628 Caruana, R., N. Karampatziakis, and A. Yessenalina. 2008. An empirical evaluation of  
629 supervised learning in high dimensions. Pages 96–103 *in* Proceedings of the 25th  
630 international conference on machine learning. ACM.

631 Clark, J. S., S. R. Carpenter, M. Barber, S. Collins, A. Dobson, J. A. Foley, D. M.  
632 Lodge, M. Pascual, R. Pielke, W. Pizer, and others. 2001. Ecological forecasts: An  
633 emerging imperative. *Science* 293:657–660.

634 Cutler, D. R., T. C. Edwards, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J.  
635 Lawler. 2007. Random forests for classification in ecology. *Ecology* 88:2783–2792.

636 Detsch, F. 2016. Gimms: Download and process gimms ndvi3g data. R package version  
637 1.0.0.

638 Dietze, M. C. 2017. Ecological forecasting. Princeton University Press.

639 Dietze, M. C., A. Fox, J. L. Betancourt, M. B. Hooten, C. S. Jarnevich, T. H. Keitt, M.  
640 Kenney, C. Laney, L. Larsen, H. W. Loescher, and others. 2016. Iterative ecological  
641 forecasting: Needs, opportunities, and challenges. *in* NEON workshop:

642 Operationalizing ecological forecasting.

643 Distler, T., J. G. Schuetz, J. Velásquez-Tibatá, and G. M. Langham. 2015. Stacked  
644 species distribution models and macroecological models provide congruent projections  
645 of avian species richness under climate change. *Journal of Biogeography* 42:976–988.

646 Díaz, S., S. Demissew, J. Carabias, C. Joly, M. Lonsdale, N. Ash, A. Larigauderie, J. R.  
647 Adhikari, S. Arico, A. Báldi, and others. 2015. The ipbes conceptual  
648 framework—connecting nature and people. *Current Opinion in Environmental*  
649 *Sustainability* 14:1–16.

650 Dornelas, M., N. J. Gotelli, B. McGill, H. Shimadzu, F. Moyes, C. Sievers, and A. E.  
651 Magurran. 2014. Assemblage time series reveal biodiversity change but not systematic  
652 loss. *Science* 344:296–299.

653 Elith, J., J. R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression  
654 trees. *Journal of Animal Ecology* 77:802–813.

655 Ernest, S. M., and J. H. Brown. 2001. Homeostasis and compensation: The role of  
656 species and resources in ecosystem stability. *Ecology* 82:2118–2132.

657 Garnier, S. 2017. viridis: Default color maps from 'matplotlib'. R package version  
658 0.4.0.

659 Harris, D. J. 2015. Generating realistic assemblages with a joint species distribution  
660 model. *Methods in Ecology and Evolution* 6:465–473.

661 Harris, D. J., E. White, and S. D. Taylor. 2017a. Weecology/bbs-forecasting. Zenodo.  
662 DOI:10.5281/zenodo.888989.

663 Harris, D. J., E. White, and S. D. Taylor. 2017b. Weecology/forecasts: V0.0.2. Zenodo.  
664 DOI:10.5281/zenodo.1101123.

665 Henry, L., and H. Wickham. 2017. purrr: Functional programming tools. R package

666 version 0.2.2.2.

667 Hijmans, R. J. 2016. raster: Geographic data analysis and modeling. R package version  
668 2.5-8.

669 Houlahan, J. E., S. T. McKinney, T. M. Anderson, and B. J. McGill. 2017. The priority  
670 of prediction in ecological understanding. *Oikos* 126:1–7.

671 Hurlbert, A. H., and J. P. Haskell. 2002. The effect of energy and seasonality on avian  
672 species richness and community composition. *The American Naturalist* 161:83–97.

673 Hurlbert, A. H., and E. P. White. 2005. Disparity between range map-and survey-based  
674 analyses of species richness: Patterns, processes and implications. *Ecology Letters*  
675 8:319–327.

676 Hutchinson, R. A., L.-P. Liu, and T. G. Dietterich. 2011. Incorporating boosted  
677 regression trees into ecological latent variable models. Pages 1343–1348 *in* Proceedings  
678 of the twenty-fifth aaai conference on artificial intelligence. San Francisco, California.

679 Hyndman, R. J. 2017. forecast: Forecasting functions for time series and linear models.  
680 R package version 8.1.

681 Hyndman, R. J., and G. Athanasopoulos. 2014. Forecasting: Principles and practice.  
682 OTexts.

683 Hyndman, R. J., and Y. Khandakar. 2008. Automatic time series forecasting: The  
684 forecast package for R. *Journal of Statistical Software* 26:1–22.

685 IPCC. 2014. Summary for policymakers. *in* C. Field, V. Barros, D. Dokken, K. Mach,  
686 M. Mastrandrea, T. Bilir, M. Chatterjee, K. Ebi, Y. Estrada, R. Genova, B. Girma, E.  
687 Kissel, A. Levy, S. MacCracken, P. Mastrandrea, and L. White, editors. Climate change  
688 2014: Impacts, adaptation, and vulnerability. Part A: Global and sectoral aspects.  
689 Contribution of Working Group II to the Fifth Assessment Report of the

690 Intergovernmental Panel on Climate Change. Cambridge University Press.

691 Jarvis, A., H. Reuter, A. Nelson, and E. Guevara. 2008. Hole-filled SRTM for the globe

692 Version 4, available from the CGIAR-CSI SRTM 90m Database.

693 Jetz, W., D. S. Wilcove, and A. P. Dobson. 2007. Projected impacts of climate and

694 land-use change on the global diversity of birds. *PLoS biology* 5:e157.

695 Jolliffe, I. T., and D. B. Stephenson, editors. 2003. Forecast verification: a practitioner's

696 guide in atmospheric science. John Wiley; Sons, Ltd.

697 Kearney, M., and W. Porter. 2009. Mechanistic niche modelling: Combining

698 physiological and spatial data to predict species' ranges. *Ecology letters* 12:334–350.

699 Kent, R., A. Bar-Massada, and Y. Carmel. 2014. Bird and mammal species composition

700 in distinct geographic regions and their relationships with environmental factors across

701 multiple spatial scales. *Ecology and evolution* 4:1963–1971.

702 La Sorte, F. A., and W. J. Boecklen. 2005. Changes in the diversity structure of avian

703 assemblages in north america. *Global Ecology and Biogeography* 14:367–378.

704 Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. *R News*

705 2:18–22.

706 Maguire, K. C., D. Nieto-Lugilde, J. L. Blois, M. C. Fitzpatrick, J. W. Williams, S.

707 Ferrier, and D. J. Lorenz. 2016. Controlled comparison of species- and

708 community-level models across novel climates and communities. *Proceedings of the*

709 *Royal Society B: Biological Sciences* 283:20152817.

710 McGill, B. J. 2012. Ecologists need to do a better job of prediction – part ii – partly

711 cloudy and a 20% chance of extinction (or the 6 p's of good prediction).

712 McGlinn, D., H. Senyondo, S. Taylor, and E. White. 2017. *rdataretriever*: R interface to

713 the data retriever. R package version 1.0.0.

714 Morris, B. D., and E. P. White. 2013. The ecodata retriever: Improving access to  
715 existing ecological data. *PLOS One* 8:e65848.

716 Oedekoven, C. S., D. A. Elston, P. J. Harrison, M. J. Brewer, S. T. Buckland, A.  
717 Johnston, S. Foster, and J. W. Pearce-Higgins. 2017. Attributing changes in the  
718 distribution of species abundance to weather variables using the example of british  
719 breeding birds. *Methods in Ecology and Evolution*.

720 Pardieck, K. L., D. J. Ziolkowski Jr, Lutmerding M, K. Campbell, and M.-A. Hudson.  
721 2017. North american breeding bird survey dataset 1966 - 2016, version 2016.0. U.S.  
722 Geological Survey, Patuxent Wildlife Research Center.

723 Pebesma, E. J., and R. S. Bivand. 2005. Classes and methods for spatial data in R. *R*  
724 *News* 5:9–13.

725 Pellissier, L., A. Espíndola, J.-N. Pradervand, A. Dubuis, J. Pottier, S. Ferrier, and A.  
726 Guisan. 2013. A probabilistic approach to niche-based community models for spatial  
727 forecasts of assemblage properties and their uncertainties. *Journal of Biogeography*  
728 40:1939–1946.

729 Perretti, C. T., S. B. Munch, and G. Sugihara. 2013. Model-free forecasting  
730 outperforms the correct mechanistic model for simulated and experimental data.  
731 *Proceedings of the National Academy of Sciences* 110:5253–5257.

732 Petchey, O. L., M. Pontarp, T. M. Massie, S. K. Efi, A. Ozgul, M. Weilenmann, G. M.  
733 Palamara, F. Altermatt, B. Matthews, J. M. Levine, D. Z. Childs, B. J. McGill, M. E.  
734 Schaepman, B. Schmid, P. Spaak, A. P. Beckerman, F. Pennekamp, and I. S. Pearse.  
735 2015. The ecological forecast horizon, and examples of its uses and determinants.  
736 *Ecology Letters* 18:597–611.

737 Pinzon, J. E., and C. J. Tucker. 2014. A non-stationary 1981–2012 avhrr ndvi3g time

738 series. *Remote Sensing* 6:6929–6960.

739 PRISM Climate Group, O. S. U. 2004. PRISM gridded climate data.  
 740 <http://prism.oregonstate.edu/>.

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

743 Rapacciuolo, G., D. B. Roy, S. Gillings, R. Fox, K. Walker, and A. Purvis. 2012.  
 744 Climatic associations of british species distributions show good transferability in time  
 745 but low predictive accuracy for range change. *PLoS One* 7:e40212.

746 Ridgeway, G., with contributions from others. 2017. *gbm: Generalized boosted*  
 747 *regression models*. R package version 2.1.3.

748 Rosenzweig, M. L. 1995. *Species diversity in space and time*. Cambridge University  
 749 Press.

750 Sauer, J. R., B. G. Peterjohn, and W. A. Link. 1994. Observer differences in the north  
 751 american breeding bird survey. *The Auk*:50–62.

752 Schurr, F. M., J. Pagel, J. S. Cabral, J. Groeneveld, O. Bykova, R. B. O’Hara, F. Hartig,  
 753 W. D. Kissling, H. P. Linder, G. F. Midgley, and others. 2012. How to understand  
 754 species’ niches and range dynamics: A demographic research agenda for biogeography.  
 755 *Journal of Biogeography* 39:2146–2162.

756 Senyondo, H., B. D. Morris, A. Goel, A. Zhang, A. Narasimha, S. Negi, D. J. Harris, D.  
 757 Gertrude Digges, K. Kumar, A. Jain, K. Pal, K. Amipara, and E. P. White. 2017.  
 758 Retriever: Data retrieval tool. *The Journal of Open Source Software* 2:451.

759 Silver, N. 2012. *The signal and the noise: Why so many predictions fail—but some don’t*.  
 760 Penguin.

761 Stan Development Team. 2016. *RStan: The R interface to Stan*. R package version

762 2.14.1.

763 Stephens, J. 2016. yaml: Methods to convert r data to yaml and back. R package  
764 version 2.1.14.

765 Tetlock, P. E., and D. Gardner. 2016. Superforecasting: The art and science of  
766 prediction. Random House.

767 Thomas, C. D., A. Cameron, R. E. Green, M. Bakkenes, L. J. Beaumont, Y. C.  
768 Collingham, B. F. Erasmus, M. F. De Siqueira, A. Grainger, L. Hannah, and others.  
769 2004. Extinction risk from climate change. *Nature* 427:145–148.

770 Thuiller, W., S. Lavergne, C. Roquet, I. Boulangeat, B. Lafourcade, and M. B. Araujo.  
771 2011. Consequences of climate change on the tree of life in europe. *Nature* 470:531.

772 Tilman, D., M. Clark, D. R. Williams, K. Kimmel, S. Polasky, and C. Packer. 2017.  
773 Future threats to biodiversity and pathways to their prevention. *Nature* 546:73–81.

774 Tredennick, A. T., M. B. Hooten, C. L. Aldridge, C. G. Homer, A. R. Kleinhesselink,  
775 and P. B. Adler. 2016. Forecasting climate change impacts on plant populations over  
776 large spatial extents. *Ecosphere* 7.

777 Urban, M. C. 2015. Accelerating extinction risk from climate change. *Science*  
778 348:571–573.

779 Vellend, M., L. Baeten, I. H. Myers-Smith, S. C. Elmendorf, R. Beauséjour, C. D.  
780 Brown, P. De Frenne, K. Verheyen, and S. Wipf. 2013. Global meta-analysis reveals no  
781 net change in local-scale plant biodiversity over time. *Proceedings of the National*  
782 *Academy of Sciences* 110:19456–19459.

783 Ward, E. J., E. E. Holmes, J. T. Thorson, and B. Collen. 2014. Complexity is costly: A  
784 meta-analysis of parametric and non-parametric methods for short-term population  
785 forecasting. *Oikos* 123:652–661.

786 Warton, D. I., F. G. Blanchet, R. B. O’Hara, O. Ovaskainen, S. Taskinen, S. C. Walker,

787 and F. K. Hui. 2015. So many variables: Joint modeling in community ecology. Trends  
788 in Ecology & Evolution 30:766–779.

789 White, E. P. 2004. Two-phase species–time relationships in north american land birds.  
790 Ecology Letters 7:329–336.

791 White, E. P. 2007. Spatiotemporal scaling of species richness: Patterns, processes, and  
792 implications. Scaling biodiversity (eds D. Storch, PA Marquet & JH Brown):325–346.

793 White, E. P., and A. H. Hurlbert. 2010. The combined influence of the local  
794 environment and regional enrichment on bird species richness. The American Naturalist  
795 175:E35–E43.

796 Wickham, H. 2009. Ggplot2: Elegant graphics for data analysis. Springer-Verlag New  
797 York.

798 Wickham, H. 2017. Tidy: Easily tidy data with 'spread()' and 'gather()' functions. R  
799 package version 0.6.3.

800 Wickham, H., R. Francois, L. Henry, and K. Müller. 2017. Dplyr: A grammar of data  
801 manipulation. R package version 0.7.1.

802 Widgren, S., and others. 2016. git2r: Provides access to git repositories. R package  
803 version 0.14.0.

804 Yu, J., W.-K. Wong, and R. A. Hutchinson. 2010. Modeling experts and novices in  
805 citizen science data for species distribution modeling. Pages 1157–1162 in Data mining  
806 (icdm), 2010 ieee 10th international conference on. IEEE.