Deciphering ecology from statistical artefacts: Competing influence of sample size, prevalence and habitat specialization on species distribution models and how small evaluation datasets can inflate metrics of performance

Tyler A. Hallman | William D. Robinson

Department of Fisheries and Wildlife, Oregon State University, Corvallis, Oregon

Correspondence
Tyler A. Hallman, Department of Fisheries and Wildlife, Oregon State University, 104 Nash Hall, Corvallis, OR 97331. Email: tyler.hallman@oregonstate.edu

Editor: Raimundo Real

Abstract

Aim: Sample size and species characteristics, including prevalence and habitat specialization, can influence the predictive performance of species distribution models (SDMs). There is little agreement, however, on which metric of model performance to use. Here, we directly compare AUC and partial ROC as metrics of SDM performance through analyses on the effects of species traits and sample size on SDM performance.

Location: Three counties dominated by agricultural lands and coniferous forest in Oregon’s Willamette Valley and Coast Range ecoregions.

Methods: We systematically reduced a large avian point count dataset to alter sample sizes of 22 species of songbird. We used boosted regression trees to run SDMs for each species, quantified habitat specialization, and used mixed effects models to compare the influence of sample size, prevalence, and habitat specialization on SDM performance, calculated as AUC and partial ROC, across species. We calculated AUC and partial ROC with subset and independent evaluation data separately to more comprehensively investigate differences in metrics.

Results: We found a positive quadratic effect of sample size and a strongly positive effect of habitat specialization on both metrics of model performance. We found a weak effect of prevalence on partial ROC and no effect in AUC. Contrary to expectations, when evaluated with a subset evaluation data, partial ROC was consistently highest in models with the smallest sample sizes. These small sample size models had correspondingly small sample sizes in subset evaluation datasets. Partial ROC evaluated with independent data and AUC evaluated with subset or independent data showed the expected positive correlation between sample size and model performance.

Main Conclusions: We found that small evaluation datasets can artificially inflate partial ROC. With literature recommended minimum SDM sample sizes as low as three, attention must be given to the effects of correspondingly low sample sizes in evaluation datasets.
INTRODUCTION

Species distribution models (SDMs) tie occurrence data to environmental covariates to predict a species’ distribution over geographic space (Araújo & Peterson, 2012). SDMs are used to address conservation and management issues from the potential impacts of climate change on the efficacy of our current reserve systems (Araújo, Cabeza, Thuiller, Hannah, & Williams, 2004; Téllez-Valdés & Díaz, 2003) to the spread of invasive species (Bae, Murphy, & García-Berthou, 2018). Given the growing use of SDMs for ecological research and conservation, developing clear guidelines for practitioners is increasingly important.

Sample size (Bean, Stafford, & Brashares, 2012), species’ prevalence (i.e. the proportion of surveyed sites occupied by a species) (Sor, Park, Boets, Goethals, & Lek, 2017) and degree of habitat specialization (Connor et al., 2018; McPherson & Jetz, 2007) are among the diverse factors that affect the predictive performance of SDMs (Proosdij, Sosef, Wieringa, & Raes, 2015; Stockwell & Peterson, 2002; Wisz et al., 2008). Generally, model performance increases with sample size, with a stronger effect when sample sizes are low (Bean et al., 2012; McPherson & Jetz, 2007; Proosdij et al., 2015). This quadratic effect of sample size on SDM performance has led to research on the minimum sufficient sample size with which to build accurate SDMs (Proosdij et al., 2015; Stockwell & Peterson, 2002; Wisz et al., 2008). Estimates of minimum sufficient sample size range from 3 (Proosdij et al., 2015) to 30 (Wisz et al., 2008) and appear to be affected by SDM algorithm (Stockwell & Peterson, 2002; Wisz et al., 2008), species prevalence and landscape complexity (Proosdij et al., 2015).

The negative relationship between species’ prevalence and SDM performance is well documented (Guo et al., 2015; Sor et al., 2017). Species’ prevalence on a landscape is driven by a combination of the degree of species’ habitat specialization and the degree of landscape heterogeneity. Prevalence and habitat specialization should be correlated. Species that can use diverse habitats are likely to be more prevalent on a landscape. When examined, species habitat specialization is positively correlated with SDM performance (Connor et al., 2018; McPherson & Jetz, 2007). Habitat specialization may be the driver behind the relationship between prevalence and SDM performance. Deciphering the interwoven influences of sample size, prevalence and habitat specialization on SDM performance may influence guidelines for effective use of SDMs.

Model evaluation is a fundamental step in producing reliable predictive SDMs. Though commonplace, area under the curve (AUC) as the standard metric for SDM evaluation may be inappropriate (Fernandes, Scherrer, & Guisan, 2018; Peterson, Papeş, & Soberón, 2008). Species with small geographic ranges tend to have higher AUC (Hernandez, Graham, Master, & Albert, 2006; Luoto, Pöyry, Heikkinen, & Saarinen, 2005; McPherson & Jetz, 2007). Although this may be related to strong habitat associations from specialization (Guo et al., 2015; Tsoar, Allouche, Steinitz, Rotem, & Kadmon, 2007), it can also be an artefact of AUC calculation. The magnitude of the difference between the extent of a species’ occurrence and the extent of the landscape being modelled greatly affects AUC (Jimenez-Valverde, 2012; Lobo, Jiménez-Valverde, & Real, 2008). Increasing the spatial extent of a study to include strongly unsuitable habitats, for example, artificially improves AUC by increasing the rate of correctly predicted absences (Beale & Lennon, 2012; Jiménez-Valverde, Lobo, & Hortal, 2008).

Partial receiver operating characteristic (partial ROC) has been introduced as an alternative for model evaluation (Peterson et al., 2008). Researchers can address many of the critiques of AUC by using partial ROC and replacing commission error on the x-axis with the proportion of area a species is predicted to be present. Furthermore, through bootstrapping and using partial ROC ratios, we can compare predictive performance to random models. AUC ranges from 0 to 1 and can be interpreted as the probability that a randomly chosen occupied site has a higher score than a randomly chosen unoccupied site. An AUC of 0.5 is a random model. Partial ROC ratios, on the other hand, range from 0 to 2 with 1 representing a random model. Partial ROC ratios can be interpreted as model accuracy compared to a random model. Though not commonly done, a similar ratio could be used for AUC. The comparative sensitivity of these two metrics of SDM performance to sample size, species prevalence and species specialization has, to our knowledge, not yet been explored.

In this study, we examine the effects of sample size, prevalence and habitat specialization on SDM discrimination, an important aspect of model performance, estimated with AUC and partial ROC. We use a machine learning algorithm to predict the distributions of 22 species of bird from datasets of systematically reduced sizes. Our primary objective is to evaluate and compare AUC and partial ROC as metrics of model performance for real species on real landscapes. We calculate AUC and partial ROC with subset and independent evaluation datasets to more comprehensively assess differences in results from the two metrics. Despite being essential to model evaluation, evaluation datasets are rarely explicitly studied. Our secondary objective is to compare the relative influence of sample size, prevalence and habitat specialization on model performance. Given the abundance of literature on the effects of sample size and prevalence on model performance, the novelty in our study stems from deciphering between the competing effects of prevalence and habitat specialization, and the comparison of AUC and partial ROC as metrics of model performance.
2 | METHODS

2.1 | Survey data, study extent and study species

We compiled point count data from the Oregon 2020 database (W. D. Robinson, T. A. Hallman, & J. R. Curtis, personal communication) from Benton, Polk and Linn counties, Oregon. Benton, Polk and Linn counties lie at the confluence of Oregon’s Willamette Valley and Coast Range ecoregions (Thorson et al., 2003). Land cover in these counties is dominated by agricultural lands in the valley and wet coniferous forests in the mountains. Patches of riparian forest, oak forest, oak savanna, and wet and upland prairie speckle the landscape. We conducted five-minute unlimited distance avian stationary surveys (i.e. point counts) from sunrise until song activity decreased about seven hours after sunrise (T. A. Hallman, & W. D. Robinson, personal communication). Surveys were spaced every 0.8 km along accessible roads and every 200 m grids within two National Wildlife Refuges. We conducted surveys during the breeding season (mid-May to early July) between 2011 and 2014. Each site was visited and surveyed once. Although Oregon 2020 surveys include the data accessible roads and every 200 m grids within two National Wildlife Refuges. We conducted surveys during the breeding season (mid-May to early July) between 2011 and 2014. Each site was visited and surveyed once. Although Oregon 2020 surveys include the data necessary for addressing imperfect detection, we used unadjusted point counts from data from the commonly used eBird database (Sullivan et al., 2009). We selected 22 passerine species representing a wide range of sample size (number of positive occurrences), prevalence (number of occurrences divided by the number of surveys) and habitat specialization within the study area (Table 1). Due to small territory size and frequent auditory cues (singing), passerines are well suited to short-duration stationary visual and auditory surveys.

2.2 | Environmental data

We compiled 25 environmental variables previously used to characterize land cover, topography and forest structure in our study area (T. A. Hallman, & W. D. Robinson, personal communication) (Table S1.1). Environmental variables were obtained in or converted to 30-m resolution rasters. Since birds can be detected far from point count locations, we used focal statistics in ArcMap to characterize environmental variables at four spatial scales: 165, 315, 615 and 1,215 m radii from cell centres. These radii span a range of scales that have previously been shown to be relevant to avian occurrence (Baladron, Isacch, Cavalli, & Bó, 2016; Boscolo & Metzger, 2009; T. A. Hallman, & W. D. Robinson, personal communication). For all land cover covariates, we calculated percentage cover within each radius. For all forest structure and topographic covariates, we calculated mean values within each radius.

2.3 | Systematic reductions in sample size

To create datasets with systematically reduced sample sizes and relatively stable prevalence for all study species, we iteratively halved the full dataset three times (Figure S1.1). Each iteration was a systematic thinning performed by removing every other count. Where the distance between roadside counts was approximately 0.8 km in the full dataset, the distance between roadside counts was around 1.6 km in the least reduced dataset, 3.2 km in the second most reduced dataset and 6.4 km in the most reduced dataset (Figure S1.2).

2.4 | Species distribution models

We used binomial boosted regression trees (BRTs) for all species distribution models (Figure S1.1). Boosted regression trees frequently perform among the best of the commonly used SDM algorithms (Elith & Graham, 2009; García-Callejas & Araújo, 2016; Guisan, Graham, Elith, & Huettemann, 2007). For all BRTs, tree complexity was 3 (Johnston et al., 2015), bag fraction was held constant at 0.75, and learning rates were optimized to allow for a minimum of 1,000 trees (Elith, Leathwick, & Hastie, 2008). BRTs were constructed through a 10-fold cross-validation method. We used a multiscale framework where environmental covariates were included at each of their four scales of characterization (T. A. Hallman, & W. D. Robinson, personal communication). Multiscale SDMs of this type consistently perform as well or better than optimized single-scale models. For each BRT, we withheld a random 20% of the dataset for model evaluation (herein: subset evaluation data). We ran ten iterations of each model to address the stochasticity inherent in the data splitting process (into training and evaluation datasets) and the BRT algorithm.

2.5 | Model evaluation

We used withheld subset evaluation data and survey data from similar habitats in an adjacent county for model evaluation (herein: independent evaluation data). Though from the same Oregon 2020 dataset, no independent evaluation data were used in model training and within a species, the same independent evaluation data were used for each iteration. We calculated AUC (Pearce & Ferrier, 2000) and partial ROC (Peterson et al., 2008) on the subset and independent evaluation datasets. These metrics measure model discrimination capacity, an important component of model performance, and are herein referred to as metrics of model performance. Partial ROC was calculated with an omission error threshold of 5% (Martínez-Gutiérrez, Martínez-Meyer, Palomares, & Fernández, 2018; Ruiz-Sánchez, Renton, Landgrave-Ramirez, Mora-Aguilar, & Rojas-Soto, 2015). To create confidence intervals and compare partial ROC from our models to random models, we resampled 50% of each dataset with replacement for 500 iterations. We used a Wilcoxon test to statistically test the difference between AUC ratios from our models to those of random models. We report partial ROC values as partial ROC ratios (compared to random models) throughout. We ran BRTs with the dismo package in R (Hijmans, Phillips, Leathwick, & Elith, 2017). We calculated Partial ROC with the niche toolbox.
### TABLE 1  Prevalence, sample size and habitat specialization of 22 passerine study species

| Species                  | Full prevalence | Mean full training sample size | Mean minimum training sample size | Mean full subset evaluation sample size | Mean minimum subset evaluation sample size | Independent evaluation sample size | CV of proportional habitat | Weighted Simpson's diversity index |
|--------------------------|-----------------|-------------------------------|-----------------------------------|-----------------------------------------|---------------------------------------------|-----------------------------------|-----------------------------|---------------------------------|
| American Robin           | 0.49            | 1,127.5                       | 149.1                             | 286.5                                   | 35.9                                        | 104.0                             | 0.248                        | 0.953                           |
| Black-capped Chickadee   | 0.19            | 442.4                         | 61.0                              | 110.6                                   | 14.0                                        | 42.0                              | 0.431                        | 0.843                           |
| Black-headed Grosbeak    | 0.35            | 811.2                         | 96.4                              | 203.8                                   | 20.6                                        | 94.0                              | 0.149                        | 0.981                           |
| Brown Creeper            | 0.09            | 215.8                         | 28.3                              | 56.2                                    | 6.7                                         | 19.0                              | 0.361                        | 0.836                           |
| Black-throated Grey Warbler | 0.14       | 324.1                         | 38.9                              | 82.9                                    | 10.1                                        | 41.0                              | 0.468                        | 0.834                           |
| Chestnut-backed Chickadee | 0.16          | 371.1                         | 48.0                              | 94.9                                    | 13.0                                        | 27.0                              | 0.575                        | 0.755                           |
| Common Yellowthroat      | 0.26            | 603.7                         | 63.2                              | 146.3                                   | 14.8                                        | 37.0                              | 0.450                        | 0.832                           |
| Hermit Warbler           | 0.22            | 521.2                         | 64.9                              | 131.8                                   | 16.1                                        | 45.0                              | 0.821                        | 0.656                           |
| House Finch              | 0.12            | 266.3                         | 46.9                              | 70.7                                    | 11.1                                        | 24.0                              | 0.689                        | 0.652                           |
| Lazuli Bunting           | 0.10            | 232.7                         | 32.8                              | 61.3                                    | 8.2                                         | 39.0                              | 0.462                        | 0.803                           |
| Orange-crowned Warbler   | 0.26            | 616.4                         | 77.4                              | 150.6                                   | 19.6                                        | 44.0                              | 0.204                        | 0.953                           |
| Pacific Wren             | 0.21            | 489.5                         | 49.0                              | 125.5                                   | 15.0                                        | 26.0                              | 0.945                        | 0.633                           |
| Pacific-slope Flycatcher | 0.30            | 697.8                         | 84.5                              | 175.2                                   | 21.5                                        | 24.0                              | 0.652                        | 0.733                           |
| Savannah Sparrow         | 0.15            | 350.0                         | 46.9                              | 87.0                                    | 11.1                                        | 43.0                              | 0.813                        | 0.526                           |
| Song Sparrow             | 0.48            | 1,109.4                       | 132.2                             | 280.6                                   | 30.8                                        | 82.0                              | 0.209                        | 0.966                           |
| Spotted Towhee           | 0.36            | 830.9                         | 100.4                             | 204.1                                   | 25.6                                        | 82.0                              | 0.230                        | 0.965                           |
| Swainson's Thrush        | 0.49            | 1,145.1                       | 139.5                             | 284.9                                   | 31.5                                        | 88.0                              | 0.371                        | 0.888                           |
| Warbling Vireo           | 0.28            | 662.2                         | 82.3                              | 163.8                                   | 21.7                                        | 33.0                              | 0.538                        | 0.819                           |
| Western Wood-Pewee       | 0.21            | 480.7                         | 55.2                              | 122.3                                   | 13.8                                        | 49.0                              | 0.397                        | 0.872                           |
| Willow Flycatcher        | 0.08            | 174.1                         | 21.8                              | 44.9                                    | 6.2                                         | 31.0                              | 0.546                        | 0.803                           |
| Wilson’s Warbler         | 0.31            | 733.9                         | 82.7                              | 178.1                                   | 20.3                                        | 25.0                              | 0.568                        | 0.796                           |
| Yellow Warbler           | 0.07            | 155.8                         | 15.1                              | 37.2                                    | 5.9                                         | 19.0                              | 0.633                        | 0.676                           |

Note: Training and evaluation sample sizes are the mean of ten iterations of the same model in the full and most reduced datasets. Habitat specialization from top models included.
(NicheToolBox, 2017). Spatial autocorrelation can inflate metrics of model performance when evaluated with randomly withheld data. Since subset AUC and partial ROC were calculated on the same subset evaluation data, they are subject to identical spatial autocorrelation, and comparison between metrics should be unaffected (the primary objective of this study). Independent AUC and partial ROC were calculated on independent data that should be free from issues of spatial autocorrelation in training and evaluation datasets.

### 2.6 Metrics of habitat specialization

To quantify habitat specialization, we calculated indices of species habitat specialization for all study species. To calculate our indices, we created a 300 m buffer around each count location. We used the tabulate area tool in ArcMap to calculate the area of each of the six most common classes from the subclass level of the National GAP Project land cover classification database. For each species, we calculated the area of all six land cover classes at sites where the species occurred. From these data, we adopted a series of indices used for resource partitioning (DeVictor et al., 2010): the Simpson’s Diversity Index, the Shannon Index, the Simpson’s Diversity Index weighted by habitat availability (herein: Weighted Simpson’s Diversity Index) and a similarly weighted index of diversity proposed by Smith (Smith, 1982). We also calculated the coefficient of variation of the proportion of available habitat occupied (herein: CV of Proportional Habitat) by the species for each of the six land cover classes. Like the weighted indices above, the CV of Proportional Habitat is inherently weighted by habitat availability and accounts for differences in overall species prevalence.

### 2.7 Mixed-effects models of prevalence, sample size and habitat specialization

A species’ true prevalence on the landscape is an unknown latent variable. In this study, we used sample prevalence as a surrogate for landscape prevalence. We calculated sample prevalence as the sample size divided by the total number of surveys within a single dataset (e.g. full or reduced datasets). Within a single dataset, species prevalence and sample size were therefore inextricably linked and their independent effects on SDM performance could not be independently evaluated. Through the systematic reductions in dataset size, however, sample size was manipulated for each species with little effect on prevalence (see below).

We used mixed-effects models to evaluate the influence of species prevalence, sample size and habitat specialization on model performance. To do so, we used data from all full- and reduced-dataset SDMs of all study species within the same mixed-effects models. We ran separate analyses with AUC and partial ROC as dependent variables. Furthermore, we ran separate analyses for model performance as estimated with subset and independent evaluation datasets. We scaled and centred all variables included in mixed-effects models. For all mixed-effects models, we treated species as a random effect. Before incorporating prevalence and sample size, we ran mixed-effects models that included each of our five indices of habitat specialization. We used the habitat specialization index included in the BIC top-performing model as the single metric of habitat specialization included in the subsequent model set. We then built a mixed-effects model set with habitat specialization, sample size and prevalence as independent variables. We also examined quadratic effects of sample size and prevalence. We included single- and additive multivariable models as well as models with an interaction effect of prevalence and sample size. We compared models through BIC. The model with the lowest BIC was considered the best model, and models within four BIC were considered competitive. We conducted this analysis with both subset evaluation data and independent evaluation data.

To better understand the relationship between AUC and partial ROC, we ran regressions between the two for metrics calculated with both the subset evaluation and the independent evaluation data. To examine consistency of test metrics based on source of evaluation data, we ran a regression of AUC calculated with the subset evaluation data on AUC calculated with the independent evaluation data. Likewise, we ran a regression of partial ROC from each of our two sources of evaluation data. We ran these linear regressions on the mean values of the ten iterations of each SDM to avoid excessive pseudoreplication. To better visualize the effects of sample size on each metric of model performance, we calculated the percentage full AUC and percentage full partial ROC from models from each of the reduced datasets. Finally, we ran regressions to examine the relationship between prevalence and habitat specialization in the full dataset.

### 2.8 Evaluation dataset size

Subset evaluation datasets decreased in size with dataset thinning. This reduction in evaluation dataset size could impact metrics of model performance. In BIC top models with nonsensical covariate values, we iteratively imposed increasing thresholds of minimum evaluation dataset size and reran top models excluding any SDMs that did not meet threshold values. We recorded covariate values with each iteration.

### 3 RESULTS

#### 3.1 Prevalence, sample size and specialization

In the full dataset, mean species prevalence ranged from 0.07 to 0.49 while mean species sample size ranged from 193 to 1,430 (Table 1). In the most reduced dataset, prevalence across species ranged from 0.06 to 0.51 while sample size ranged from 21 to 185. The mean within species standard deviation of prevalence across full and reduced datasets was 0.009, indicating negligible changes
in prevalence with reductions in dataset size. Species habitat specialization ranged from highly generalist to highly specialist values in each of the indices (Tables 1, S1.2).

### 3.2 | AUC

Across datasets, mean AUCs of SDMs for a species calculated with the subset evaluation data ranged from 0.67 to 0.93 with an across species mean of 0.81 (Table S1.3). Within habitat specialization mixed-effects models, the model with CV of Proportional Habitat was over four BIC better than the next best model and had over 83% of the model weight (Table S1.4). In the subsequent mixed-effects models comparing the effects of prevalence, sample size and habitat specialization, the model including specialization and a quadratic effect of occurrence was over four BIC better than the next best model and had over 86% of the model weight (Table 3). In this BIC top model, there was a positive effect of sample size (slope = 0.105, p-value < .0001) with a diminishing effect at larger sample sizes (slope = −0.033, p-value < .0001) and a positive effect of CV of Proportional Habitat (slope = 3.157, p-value < .0004). Given the extrapolation of the model to a neighbouring independent evaluation area, weaker relationships and lower model performance are to be expected.

### 3.3 | Partial ROC

Across species, partial ROCs of SDMs calculated with the subset evaluation data ranged from 1.05 to 1.58 with a mean of 1.27 (Table S1.3). Within habitat specialization mixed-effects models, the model with Weighted Simpson’s Diversity Index was over two BIC better than the next best model and had over 77% of the model weight (Table S1.5). In the subsequent mixed-effects models comparing the effects of prevalence, sample size and habitat specialization, the model including specialization, prevalence and a quadratic effect of occurrence was over four BIC better than the next best model and had over 88% of the model weight (Table 4). In this BIC top model, there was a negative effect of sample size (slope = −0.228, p-value < .0001) with an increasing effect at larger sample sizes (slope = 0.071, p-value < .0001), a positive effect of prevalence (slope = 0.586, p-value = .0196) and a negative effect of Weighted Simpson’s Diversity Index (slope = −7.756, p-value < .0001). Given the extrapolation of the model to a neighbouring independent evaluation area, weaker relationships and lower model performance are to be expected.

**Table 2**  BIC mixed-effects model set comparing the effects of habitat specialization, sample size and prevalence on AUC of species distribution models from 22 study species calculated with subset evaluation data

| Model                                      | df | LL    | BIC     | Delta BIC | Model weight |
|--------------------------------------------|----|-------|---------|-----------|--------------|
| Specialization + Quadratic Occurrence      | 6  | 734.49| 1,509.66| 0.00      | 0.96         |
| Specialization + Quadratic Occurrence + Prevalence | 7  | 734.40| 1,516.26| 6.60      | 0.04         |
| Specialization + Occurrence × Prevalence   | 7  | 739.85| 1,527.15| 17.49     | 0.00         |
| Specialization + Occurrence                | 5  | 749.07| 1,532.04| 22.38     | 0.00         |
| Specialization + Occurrence + Prevalence   | 6  | 748.65| 1,537.98| 28.32     | 0.00         |
| Quadratic Occurrence                       | 5  | 753.68| 1,541.27| 31.60     | 0.00         |
| Quadratic Occurrence + Prevalence          | 6  | 751.62| 1,543.93| 34.26     | 0.00         |
| Specialization + Occurrence + Quadratic Prevalence | 7  | 748.51| 1,544.49| 34.82     | 0.00         |
| Occurrence + Prevalence                    | 6  | 757.61| 1,555.89| 46.23     | 0.00         |
| Occurrence                                 | 4  | 768.13| 1,563.39| 53.73     | 0.00         |
| Occurrence + Prevalence                    | 5  | 767.20| 1,568.29| 58.63     | 0.00         |
| Occurrence + Quadratic Prevalence          | 6  | 766.76| 1,574.19| 64.53     | 0.00         |
| Specialization                             | 4  | 782.07| 1,591.27| 81.60     | 0.00         |
| Specialization + Prevalence                | 5  | 778.73| 1,591.36| 81.69     | 0.00         |
| Specialization + Quadratic Prevalence      | 6  | 778.70| 1,598.07| 88.41     | 0.00         |
| Null                                       | 3  | 798.82| 1,617.98| 108.31    | 0.00         |
| Prevalence                                 | 4  | 798.80| 1,624.72| 115.05    | 0.00         |
| Quadratic Prevalence                       | 5  | 798.69| 1,631.27| 121.61    | 0.00         |

Note: Species was treated as a random effect throughout models.
p-value = .0075). Weighted Simpson’s Diversity Index is an index with lower values indicating greater habitat specialization. The negative slope of the effect on partial ROC therefore indicates greater model performance with greater habitat specialization. Aside from habitat specialization, these coefficients contradict those found for AUC.

Patterns in partial ROC calculated with the independent evaluation data starkly contrasted partial ROC calculated with subset evaluation data. Across species, partial ROCs ranged from 1.08 to 1.52 with a mean of 1.25 (Table S1.3). In the mixed-effects models comparing the effects of prevalence, sample size and habitat specialization, the model including specialization, prevalence and a quadratic effect of occurrence was over two BIC better than the next best model and had over 59% of the model weight (Table 5). In the BIC top model, there was a positive effect of sample size (slope = 0.172, p-value < .0001) with a diminishing effect at larger sample sizes (slope = −0.042, p-value < .0001), a negative effect of prevalence (slope = −0.291, p-value = .707) and a negative effect of Weighted Simpson’s Diversity Index (slope = −4.579, p-value = .0003). Models containing a quadratic effect of sample size on partial ROC combined for over 98% of the total model weight. In each of the top three models, which combined for over 97% of the total model weight, coefficients for prevalence, specialization and quadratic sample size were similar.

3.4 | AUC and partial ROC comparison

There was a strong positive correlation between AUCs calculated with subset and independent evaluation data ($R^2 = .501$, p-value < .0001; Figure 1a). The correlation between partial ROCs calculated with subset and independent evaluation data was much weaker ($R^2 = .056$, p-value = .149; Figure 1b). There was a weak correlation between AUCs and partial ROCs calculated with the subset evaluation data ($R^2 = .151$, p-value = .0001; Figure 2). In contrast, there was a strong correlation between AUCs and partial ROCs calculated with the independent evaluation data ($R^2 = .790$, p-value < .0001). There was a significant correlation between CV of Proportional Habitat and prevalence ($R^2 = .212$, p-value = .0179; Figure 3a), and between Weighted Simpson’s Diversity Index and prevalence ($R^2 = .316$, p-value = .0038; Figure 3b). This relationship was nearly identical when model performance was calculated with subset or independent evaluation data.

3.5 | Evaluation dataset size

We examined evaluation dataset size as an explanation for the inconsistency of the results in partial ROC calculated with the subset evaluation data compared to all other results in this study. In the full dataset, evaluation dataset size in the subset evaluation data ranged from 26 to 304. In the most reduced dataset, evaluation dataset size in the subset evaluation data ranged from 5 to 40. Evaluation dataset size in the independent evaluation data, which was not subject to reductions in data set size, ranged from 19 to 104. When iteratively imposing an increasing minimum threshold on evaluation dataset size from 5 to 150, the coefficients in the BIC top-performing model of

| Model | df | LL | BIC | Delta BIC | Model weight |
|-------|----|----|-----|-----------|--------------|
| Specialization + Quadratic Occurrence | 6 | −201.88 | 444.45 | 0.00 | 0.87 |
| Specialization + Occurrence × Prevalence | 7 | −200.97 | 449.40 | 4.95 | 0.07 |
| Specialization + Quadratic Occurrence + Prevalence | 7 | −201.77 | 451.00 | 6.55 | 0.03 |
| Quadratic Occurrence | 5 | −209.08 | 452.06 | 7.61 | 0.02 |
| Specialization + Occurrence | 5 | −210.78 | 455.46 | 11.01 | 0.00 |
| Occurrence × Prevalence | 6 | −208.33 | 457.33 | 12.88 | 0.00 |
| Quadratic Occurrence + Prevalence | 6 | −208.91 | 458.50 | 14.05 | 0.00 |
| Specialization + Occurrence + Prevalence | 6 | −210.18 | 461.04 | 16.59 | 0.00 |
| Occurrence | 4 | −217.92 | 462.93 | 18.51 | 0.00 |
| Specialization + Occurrence + Quadratic Prevalence | 7 | −208.31 | 464.09 | 19.64 | 0.00 |
| Specialization | 4 | −219.24 | 465.59 | 21.14 | 0.00 |
| Specialization + Prevalence | 5 | −217.31 | 468.52 | 24.07 | 0.00 |
| Occurrence + Prevalence | 5 | −217.92 | 469.73 | 25.29 | 0.00 |
| Occurrence + Quadratic Prevalence | 6 | −215.65 | 471.99 | 27.54 | 0.00 |
| Null | 3 | −226.10 | 472.53 | 28.08 | 0.00 |
| Specialization + Quadratic Prevalence | 6 | −215.97 | 472.62 | 28.18 | 0.00 |
| Prevalence | 4 | −225.64 | 478.40 | 33.95 | 0.00 |
| Quadratic Prevalence | 5 | −224.03 | 481.95 | 37.50 | 0.00 |

Note: Species was treated as a random effect throughout models.
partial ROC calculated with the subset evaluation data moved towards expected values that match the other results in this study: the effect of sample size became positive (Figure 4a), the quadratic term for sample size became negative (Figure 4b), and the effect of prevalence became negative (Figure 4c). The effect of habitat specialization decreased, which may be due to

### TABLE 4
BIC mixed-effects model set comparing the effects of habitat specialization, sample size and prevalence on partial ROC of species distribution models from 22 study species calculated with subset evaluation data

| Model                                    | df | LL   | BIC  | Delta BIC | Model weight |
|------------------------------------------|----|------|------|-----------|--------------|
| Specialization + Quadratic Occurrence + Prevalence | 7  | -545.47 | 1,138.40 | 0.00 | 0.89 |
| Specialization + Quadratic Occurrence    | 6  | -551.46 | 1,143.60 | 5.20 | 0.07 |
| Specialization + Occurrence × Prevalence  | 7  | -548.68 | 1,144.83 | 6.43 | 0.04 |
| Quadratic Occurrence + Prevalence        | 6  | -553.43 | 1,147.54 | 9.14 | 0.01 |
| Quadratic Occurrence                     | 5  | -558.75 | 1,151.40 | 13.00 | 0.00 |
| Occurrence × Prevalence                  | 6  | -556.79 | 1,154.26 | 15.86 | 0.00 |
| Specialization + Occurrence               | 5  | -566.35 | 1,166.61 | 28.21 | 0.00 |
| Specialization + Occurrence + Prevalence  | 6  | -563.45 | 1,167.57 | 29.17 | 0.00 |
| Specialization + Occurrence + Quadratic Prevalence | 7  | -562.27 | 1,171.99 | 33.59 | 0.00 |
| Occurrence                                | 4  | -573.49 | 1,174.10 | 35.70 | 0.00 |
| Occurrence + Prevalence                   | 5  | -571.36 | 1,176.62 | 38.22 | 0.00 |
| Occurrence + Quadratic Prevalence         | 6  | -569.33 | 1,179.35 | 40.95 | 0.00 |
| Specialization                            | 4  | -579.03 | 1,185.19 | 46.79 | 0.00 |
| Specialization + Prevalence               | 5  | -578.43 | 1,190.77 | 52.37 | 0.00 |
| Null                                     | 3  | -586.38 | 1,193.10 | 60.70 | 0.00 |
| Specialization + Quadratic Prevalence     | 6  | -576.95 | 1,194.57 | 56.17 | 0.00 |
| Prevalence                                | 4  | -586.17 | 1,199.47 | 61.07 | 0.00 |
| Quadratic Prevalence                      | 5  | -583.53 | 1,200.95 | 62.55 | 0.00 |

Note: Species was treated as a random effect throughout models.

### TABLE 5
BIC mixed-effects model set comparing the effects of habitat specialization, sample size and prevalence on partial ROC of species distribution models from 22 study species calculated with independent evaluation data

| Model                                    | df | LL   | BIC  | Delta BIC | Model weight |
|------------------------------------------|----|------|------|-----------|--------------|
| Specialization + Quadratic Occurrence + Prevalence | 7  | -169.35 | 386.16 | 0.00 | 0.60 |
| Specialization + Quadratic Occurrence    | 6  | -173.84 | 388.36 | 2.20 | 0.20 |
| Quadratic Occurrence + Prevalence        | 6  | -173.94 | 388.55 | 2.39 | 0.18 |
| Specialization + Occurrence × Prevalence  | 7  | -172.92 | 393.30 | 7.14 | 0.02 |
| Occurrence × Prevalence                  | 6  | -177.97 | 396.63 | 10.47 | 0.00 |
| Quadratic Occurrence                     | 5  | -182.54 | 398.98 | 12.83 | 0.00 |
| Specialization + Occurrence               | 5  | -187.15 | 408.20 | 22.04 | 0.00 |
| Specialization + Occurrence + Prevalence  | 7  | -181.00 | 409.46 | 23.30 | 0.00 |
| Specialization + Occurrence + Quadratic Prevalence | 6  | -184.88 | 410.45 | 24.29 | 0.00 |
| Occurrence + Quadratic Prevalence         | 6  | -186.24 | 413.17 | 27.01 | 0.00 |
| Occurrence + Prevalence                   | 5  | -190.38 | 414.66 | 28.50 | 0.00 |
| Occurrence                                | 4  | -195.70 | 418.53 | 32.37 | 0.00 |
| Specialization                            | 4  | -218.05 | 463.22 | 77.06 | 0.00 |
| Occurrence + Quadratic Prevalence         | 6  | -212.91 | 466.49 | 80.34 | 0.00 |
| Specialization + Prevalence               | 5  | -218.02 | 469.93 | 83.77 | 0.00 |
| Null                                     | 3  | -226.08 | 472.49 | 86.34 | 0.00 |
| Quadratic Prevalence                      | 5  | -219.37 | 472.64 | 86.48 | 0.00 |
| Prevalence                                | 4  | -225.09 | 477.31 | 91.15 | 0.00 |

Note: Species was treated as a random effect throughout models.
4 | DISCUSSION

We used a rigorous study design to compare AUC and partial ROC as metrics of SDM performance. We used subset and independent evaluation data to investigate the impacts of evaluation dataset on both metrics. Although our primary objective was to compare metrics of SDM performance, we did so by examining the comparative influence of prevalence, sample size and habitat specialization on performance estimated by each of these metrics. We used 22 species of bird that reflected a range of prevalence, habitat specialization and sample size in our analyses. Birds are "easily" detected during the breeding season, but reflect the messy reality of biodiversity surveys; many individuals and species go undetected even when present. We see this noise as an essential element in studying the effects of species and dataset characteristics on SDM performance.

4.1 | Model performance measured by AUC

Model performance for AUC compared well with literature reported values (Illán et al., 2014; Shirley et al., 2013). We found good agreement on the effects of sample size and habitat specialization on model performance measured by AUC calculated with subset and independent evaluation data. Similar to past studies increases in sample size increased model performance with a stronger effect at small sample sizes (Bean et al., 2012; McPherson & Jetz, 2007; Proosdij et al., 2015) (Figure 5a,b). Model performance was also higher in more specialized species (Connor et al., 2018; McPherson & Jetz, 2007; Soultan & Safi, 2017) (Figure 6a). Overall model performance and the strength of each effect were less pronounced in the independent evaluation data, where predictions are an extrapolation to a neighbouring landscape. These extrapolations, however, were geographic in nature, with both landscapes occupying similar environmental space (Franklin, 2013).

Prevalence appeared to have little effect on model performance as measured by AUC. As landscape prevalence is an unknown latent
variable, we used sample prevalence as a surrogate measure. Given that our surveys were conducted periodically throughout the landscape, it is reasonable to assume that sample prevalence would be positively correlated with landscape prevalence. That said, we conducted more counts in targeted rarer habitats within national wildlife refuges. These surveys could add noise to the sample prevalence–landscape prevalence relationship and could ultimately diminish the perceived effect of prevalence on model performance in our models (Meynard & Kaplan, 2012).

4.2 Model performance measured by partial ROC

Model performance for partial ROC compared well with literature reported values (Adelino, Anjos, & Lima, 2017; Martínez-Gutiérrez et al., 2018; Ruiz-Sánchez et al., 2015; Yañez-Arenas et al., 2018). Calculated with the independent evaluation data, the effects of sample size (Figure 5c,d) and habitat specialization (Figure 6b) on partial ROC matched results for AUC and literature expectations (Bean et al., 2012; Soltan & Safi, 2017). Unlike results for AUC, there was also a weak negative effect of prevalence on partial ROC. There may be two explanations for the weak strength of the effect of prevalence on model performance compared to the literature (Guo et al., 2015; Sor et al., 2017). First, as discussed above, our substitution of sample prevalence for landscape prevalence may have introduced noise that obscured the effect. As we used real species on real landscapes, we do not know true landscape prevalence. Second, habitat specialization and prevalence are correlated (Figure 3a,b). Prevalence may be a weak surrogate for habitat specialization. Habitat specialization combined with the specific composition of a landscape drives landscape prevalence. On heterogeneous landscapes, generalists are expected to have higher landscape prevalence as they can occupy many diverse habitats. On more homogenous landscapes, the relationship between habitat specialization and landscape prevalence may break down as specialists of the dominant habitat can display high prevalence. In our study, the poor performance of mixed-effects models that include prevalence without specialization indicates either that sample prevalence was not strongly correlated to landscape prevalence or that prevalence is a poor surrogate for specialization.

Calculated with the subset evaluation data, the effects of prevalence and sample size on model performance measured by partial ROC strongly contradicted our results for partial ROC calculated with independent evaluation data, our results for AUC and our expectations from the literature (Bean et al., 2012; Guo et al., 2015; McPherson & Jetz, 2007; Proosdij et al., 2015; Sor et al., 2017). Model performance decreased with increasing sample size and increased with increasing prevalence (Figure 5c,d). When landscape size and composition are kept constant, increasing sample size should not have a negative effect on model performance. The disparity between partial ROC evaluated with subset and independent data is further exemplified by their weak relationship (Figure 1b) and the weak relationship between AUC and partial ROC calculated with the subset evaluation data (Figure 2). Why then were the coefficients in the BIC top model for partial ROC calculated with the subset evaluation data contrary to expectations? The answer lies in sample size, but not the sample size normally assessed in studies of model performance.

Evaluation datasets are rarely discussed in species distribution modelling. It is common practice to withhold a percentage of the full dataset to evaluate models. By separating out data that were never used in model training, a degree of independence is maintained in model evaluation. In this study, we withheld 20% of each dataset with which to calculate the subset evaluation metrics of model performance. The answer lies in sample size, but not the sample size normally assessed in studies of model performance.

Evaluation datasets are rarely discussed in species distribution modelling. It is common practice to withhold a percentage of the full dataset to evaluate models. By separating out data that were never used in model training, a degree of independence is maintained in model evaluation. In this study, we withheld 20% of each dataset with which to calculate the subset evaluation metrics of model performance. The answer lies in sample size, but not the sample size normally assessed in studies of model performance.

Calculated with the subset evaluation data, the effects of prevalence and sample size on model performance measured by partial ROC strongly contradicted our results for partial ROC calculated with independent evaluation data, our results for AUC and our expectations from the literature (Bean et al., 2012; Guo et al., 2015; McPherson & Jetz, 2007; Proosdij et al., 2015; Sor et al., 2017). Model performance decreased with increasing sample size and increased with increasing prevalence (Figure 5c,d). When landscape size and composition are kept constant, increasing sample size should not have a negative effect on model performance. The disparity between partial ROC evaluated with subset and independent data is further exemplified by their weak relationship (Figure 1b) and the weak relationship between AUC and partial ROC calculated with the subset evaluation data (Figure 2). Why then were the coefficients in the BIC top model for partial ROC calculated with the subset evaluation data contrary to expectations? The answer lies in sample size, but not the sample size normally assessed in studies of model performance.

Evaluation datasets are rarely discussed in species distribution modelling. It is common practice to withhold a percentage of the full dataset to evaluate models. By separating out data that were never used in model training, a degree of independence is maintained in model evaluation. In this study, we withheld 20% of each dataset with which to calculate the subset evaluation metrics of model performance. The answer lies in sample size, but not the sample size normally assessed in studies of model performance.
when datasets were split into training and evaluation datasets. In some cases, evaluation datasets in this study had as few as five occurrences. This may seem unreasonably small, but when studies suggest that SDMs can be run on data with as few as three to ten occurrences (Hernandez et al., 2006; Proosdij et al., 2015; Stockwell & Peterson, 2002), the question becomes “what size would the related evaluation data be?” As we iteratively increased the minimum threshold for sample size in the subset evaluation datasets, coefficient values for prevalence, specialization and sample size in our top model moved towards values from independent evaluation data (Figure 4). The expected negative effect of prevalence on partial ROC was not consistently apparent until a minimum sample size of about 70 (Figure 4c). The expected positive effect of sample size on partial ROC occurred at a minimum threshold of 88 occurrences (Figure 4a). Both coefficients, however, were appreciably closer to expected values at minimum thresholds closer to 100.

Evaluation datasets with minimum occurrences of 100 may be prohibitively difficult to find for rarer species with small sample sizes. Independent evaluation data would be valuable, but if such data existed, it would be better used in the model training than independent
based on how partial ROC is calculated, we believe they are highly
generalizable to other taxonomic groups and geographic locations.
That said, more research is needed before the generalizability of our
results can truly be assessed.

5 | CONCLUSIONS

Much research has been conducted to evaluate the effect of
sample size on the performance of SDMs (Bean et al., 2012;
A. Jiménez-Valverde, Lobo, & Hortal, 2009; Mitchell, Monk, &
Laurenson, 2016). Due to the cost of surveys and the importance
of SDMs to conservation of rare species, identifying minimum
sufficient sample sizes for SDMs has been of particular interest
(Proosdij et al., 2015; Stockwell & Peterson, 2002; Wisz et al.,
2008). Few studies, however, specifically investigate the effect of
evaluation dataset sample size on SDM assessment. Though ac-
curate SDMs can be created with small sample sizes in the train-
ing dataset, small sample sizes in the evaluation dataset can lead
to inaccurate assessment of model performance. Specifically, we
found that small sample sizes in the evaluation dataset can inflate
partial ROC. At small evaluation dataset sample sizes, partial ROC
may therefore be an inappropriate metric of model performance.
AUC appears to be less sensitive to sample size in evaluation
datasets, though as noted elsewhere, low sensitivity does not al-
ways indicate greater accuracy (Fernandes et al., 2018). Without
critical inspection of performance metrics, studies with small
evaluation datasets should be wary of seemingly high-performing
SDMs.

ACKNOWLEDGEMENTS

We appreciate assistance with surveys from R. Moore, database
work and maintenance from R. DeMoss and data entry from the
Robinson Lab. The Robinson graduate laboratory provided helpful
comments and discussion. The work was supported by the gener-
ous endowment of the Bob and Phyllis Mace Watchable Wildlife
Professorship.

DATA AVAILABILITY STATEMENT

All survey data are available in the eBird database (eBird.org).

ORCID

Tyler A. Hallman https://orcid.org/0000-0003-2604-0548
William D. Robinson https://orcid.org/0000-0003-2240-0606

REFERENCES

Adelino, J. R. P., dos Anjos, L., & Lima, M. R. (2017). Invasive potential
of the pied crow (Corvus albus) in eastern Brazil: Best to eradicate
before it spreads. Perspectives in Ecology and Conservation, 15(3),
227–233. https://doi.org/10.1016/j.pecon.2017.07.001
Araújo, M. B., Cabeza, M., Thuiller, W., Hannah, L., & Williams, P. H.
(2004). Would climate change drive species out of reserves? An assess-
ment of existing reserve-selection methods. Global Change Biology, 10(9),
1618–1626. https://doi.org/10.1111/j.1365-2486.2004.00828.x
BIOSKETCHES

Tyler A. Hallman is a landscape ecologist with a focus on avian ecology. He studies patterns and drivers of species distributions and abundances. He is particularly interested in predicting change through time and using citizen science in distribution and abundance modelling.

W. Douglas Robinson is Mace Professor of Watchable Wildlife in the Department of Fisheries and Wildlife at Oregon State University. He studies avian ecology with an emphasis on how birds respond to forms of environmental change, including fragmentation. His group focuses on legacy data acquisition and analysis to benchmark biodiversity in this era of rapid change.

Author Contributions: T.A.H. conceived the main idea, ran analyses, and wrote the manuscript with regular input from W.D.R. W.D.R. also collected nearly all survey data used. All authors made substantial contributions to the interpretation of results and editing of the manuscript.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Hallman TA, Robinson WD. Deciphering ecology from statistical artefacts: Competing influence of sample size, prevalence and habitat specialization on species distribution models and how small evaluation datasets can inflate metrics of performance. Divers Distrib. 2020;26:315–328. https://doi.org/10.1111/ddi.13030