Data quantity is more important than its spatial bias for predictive species distribution modelling

Willson Gaul[1*], Dinara Sadykova[2], Hannah J. White[1], Lupe León-Sánchez[2], Paul Caplat[2], Mark C. Emmerson[2], Jon M. Yearsley[1]

[1] School of Biology and Environmental Science, University College Dublin, Dublin, Ireland

[2] School of Biological Sciences, The Queen's University Belfast, Belfast, United Kingdom

Corresponding Author:

Willson Gaul[1]

Email address: willson.gaul@ucdconnect.ie

1   **ABSTRACT**

2   Biological records are often the data of choice for training predictive species distribution

3   models (SDMs), but spatial sampling bias is pervasive in biological records data at multiple

4   spatial scales and is thought to impair the performance of SDMs. We simulated presences and

5   absences of virtual species as well as the process of recording these species to evaluate the effect

6   on species distribution model prediction performance of 1) spatial bias in training data, 2) sample

7   size (the average number of observations per species), and 3) the choice of species distribution

8   modelling method. Our approach is novel in quantifying and applying real-world spatial

9   sampling biases to simulated data. Spatial bias in training data decreased species distribution

10   model prediction performance, but sample size and the choice of modelling method were more

11   important than spatial bias in determining the prediction performance of species distribution

12   models.

14   **1   INTRODUCTION**

15   Biological records data ("what, where, when" records of species identity, location, and date of

16   observation) often contain large amounts of data about species occurrences over large spatial

17   areas (Isaac & Pocock, 2015). Knowing the geographic areas occupied by species is important

18   for practical and fundamental research in a variety of disciplines. Epidemiologists use maps of

19   predicted wildlife distributions to identify areas at high risk for wildlife-human transmission

20   (Deka & Morshed, 2018; Redding et al., 2019). Land managers can use knowledge of species

21   distributions in spatial planning to minimize impacts on wildlife of new infrastructure (Dyer et

22   al. 2017; Newson et al., 2017). Because complete population censuses are not available for most

23   species, species distribution models (SDMs) are often used to predict distributions of species

24   using relatively sparse observations of species. Species observation data used to train SDMs

25   must represent the study area, but when studies focus on scales of thousands (or tens- or

26   hundreds of thousands) of square kilometers, it is difficult and often expensive to collect

27   adequate data across the entire study extent. Spatially random or stratified sampling of species

28   across large spatial areas is possible, and such surveys exist for some taxa including butterflies

29   and birds (Uzarski et al., 2017), but such data are uncommon for most taxonomic groups (Isaac,

30   van Strien, August, de Zeeuw, & Roy, 2014). More commonly, data are either spatially extensive

31   but collected opportunistically (Amano, Lamming, & Sutherland, 2016), or are collected

32   according to structured study designs but are more spatially limited.

33   Collecting biological records data is relatively cheap compared to collecting data directly as

34   part of a research project (or at least the costs of collecting biological records are borne in large

35   part by individual observers rather than by data analysts) (Carvell et al., 2016). However, there is

36   an associated challenge because the analyst lacks control over where, when, and how data were

37   collected. Many biases have been documented in biological records data, including temporal,

38   spatial, and taxonomic biases (Boakes et al., 2010). Spatial sampling bias, in which some areas

39   are sampled preferentially, is particularly pervasive at all scales and across taxonomic groups

40   (Amano & Sutherland, 2013; Oliveira et al., 2016). Despite these biases, biological records are

41   often used in species distribution models. Given the ubiquitous presence of spatial sampling bias

42   in biological records data, it is important to know whether spatial bias in training data impedes

43   the ability of SDMs to correctly model species distributions. Data collection efforts often face a

44   practical trade-off between maximizing the overall quantity and the spatial evenness of new

45   records. It would thus be useful to know whether the value of biological records for SDMs can

46   best be improved by increasing the spatial evenness of recording (perhaps at the cost of the

47   overall amount of new data that is added), or by increasing the overall amount of recording (even

48   if new records are spatially biased).

49   Spatial sampling bias in biological records has similarities with sampling biases that have been

50   investigated in other settings. The field of econometrics uses the term "sample selection bias" to

51   refer to non-random sampling and has developed theory about when sampling bias is likely to

52   bias analyses (Wooldridge, 2009). A key consideration in econometrics' evaluations of sample

53   selection bias is determining whether the inclusion of data in the sample depends on predictor

54   variables that are included in the model ("exogenous" sample selection), or depends on the value

55   of the response variable ("endogenous" sample selection), or both (Wooldridge, 2009). In

56   ecology, Nakagawa (2015) similarly provides guidelines for assessing missing data in terms of

57   whether data is missing randomly or systematically with respect to other variables (see also

58   Gelman & Hill, 2006). In a machine learning context, Fan, Davidson, Zadrozny, & Yu (2005)

59   investigated the effect on predictive models of sample selection bias in which sampling is

60   associated with predictor variables - "exogenous sample selection" in the terms of Wooldridge

61   (2009) and "missing at random" in the terms of Nakagawa (2015) - and determined that most

62   predictive models could be sensitive or insensitive to sampling bias depending on particular

63   details of the dataset.

64   Biological records may have been collected with spatial sampling biases that are exogenous,

65   endogenous, or both, and datasets may contain a mix of records collected with different types of

66   bias. For example, when sampling intensity depends on proximity to roads (Oliveira et al., 2016),

67   the sampling bias is exogenous because records arise from biased sampling that depends on an

68   aspect of environmental space that can be included in models as a predictor variable. However,

69   when a birder, for example, submits a record of an unusual bird from a location where they

70   would not otherwise have submitted records, the bias is endogenous because the sampling

71   location depends on the value of the response variable (species presence). Most sampling biases

72   occur on a continuum and are not unequivocally categorizable using any existing scheme

73   (Nakagawa, 2015), making it difficult to describe exactly the biases in data or predict their effect

74   on model performance.

75   Studies testing the impact of spatially biased training data on predictive SDMs have shown

76   mixed results (Edwards, Cutler, Zimmerman, Geiser, & Moisen, 2006; Phillips et al., 2009;

77   Barbet-Massin, Jiguet, Albert, & Thuiller, 2012; Stolar & Nielsen, 2015). Phillips et al. (2009)

78   found that spatial bias in the presence records strongly reduced model performance when using a

79   pseudo-absence approach but not when using a presence-absence approach. Using a virtual

80   ecologist simulation approach (Zurell et al., 2010), Thibaud, Petitpierre, Broennimann, Davison,

81   & Guisan (2014) found that the effect of spatial sampling bias on SDM prediction performance

82   depended on the SDM modelling method, and that the effect of spatial sampling bias was smaller

83   than the effect of other factors, including sample size and choice of modelling method. Warton,

84   Renner, & Ramp (2013) provided a method for correcting for spatially biased data when building

85   SDMs, but found that the resulting improvement in model predictive performance was small.

86   Because there is no clear guidance about when spatial bias in training data will or will not affect

87   model predictions, tests of the observed effect of spatial biases common in biological records are

88   important for determining whether those biases are likely to be problematic in practice.

89   The effect of spatial sampling bias on model predictions can be studied using real or simulated

90   data (Zurell et al., 2010; Meynard, Leroy & Kaplan, 2019). Using real data has the advantage

91   that the biases in the data are, well, real. The spatial pattern, intensity, and correlation of

92   sampling bias with environmental space are exactly of the type that analyses of real data must

93   cope with. However, using real data has two disadvantages. First, the truth about the outcome

94   being modeled (species presence or absence) is not completely known in the real world, making

95   it impossible to evaluate how well models represent the truth. Second, biases in real data are not

96   limited to the biases under study – a study investigating the effect of exogenous spatial sampling

97   bias will be unable to exclude from a real dataset records generated by endogenously biased

98   sampling that depends on the values of the outcome variable. Simulation studies avoid both these

99   problems. Because the investigator specifies the underlying pattern that is subsequently modeled,

100   the truth is known exactly (even when realized instances of the simulation are generated with

101   some stochasticity). The investigator also has direct control over which biases are introduced into

102   a simulated dataset, and therefore can be more confident that any observed effects on predictions

103   are due to the biases under investigation.

104   Spatial sampling bias can be introduced into simulated data using a parametric function that

105   describes the bias (Isaac et al., 2014; Stolar & Nielsen, 2015; Thibaud et al., 2014; Simmonds,

106   Jarvis, Henrys, Isaac & O'Hara, 2020) or by following a simplified ad-hoc rule (e.g. splitting the

107   study region into distinct areas that are sampled with different intensities) (Phillips et al., 2009).

108   However, these approaches may not adequately test the effect of spatial bias if the biases found

109   in real biological records do not follow parametric functions or are more severe than artificial

110   parametric or ad-hoc biases. We used observed sampling patterns from Irish biological records to

111   sample simulated species distributions using realistic spatially biased sampling.

112   We used a virtual ecologist approach (Zurell et al., 2010; Meynard, Leroy & Kaplan, 2019)

113   applied at the scale of Ireland to investigate the effect on the predictive performance of SDMs of

114   1) spatial sampling bias, 2) sample size (the average number of records per species), and 3)

115   choice of SDM method. We quantified the spatial sampling biases used in our study to enable

116   comparison with biases in other datasets. To the best of our knowledge, our approach is novel in

117   applying real-world spatial sampling biases, derived directly from spatial sampling patterns in

118   existing datasets, to simulated virtual species.

2   METHODS

120   We assessed the ability of species distribution models to predict "virtual species" distributions

121   (Leroy, Meynard, Bellard, & Courchamp, 2016; Zurell et al., 2010) when the models were

122   trained with datasets with a range of spatial sampling biases and sample sizes. The simulation

123   and analysis process is illustrated in Fig. 1. Virtual species distributions were produced by

124   defining the responses of virtual species to environmental predictor variables (Table 1, section

125   2.1) using the approach described in section 2.4.1. We then created maps of "true" virtual

126   species distributions covering 840 10 km x 10 km grid squares in Ireland (total area of study

127   extent = 84,000 km^2).   We then generated "virtual biological records" by sampling presence-

128   only records from the community of virtual species in each grid square, using sampling patterns

129   taken from real Irish biological records data. We then inferred non-detections for each species

130   using presence records of other species (van Strien, van Swaay, & Termaat, 2013). SDMs were

131   trained using the environmental variables as predictors and the virtual species detection/non-

132   detection data as the response. Model prediction performance was evaluated using three

133   measures of prediction performance (see section 2.5).

2.1   Environmental predictor variables

135   We chose environmental predictor variables with a range of spatial patterns and scales of

136   spatial auto-correlation (Table 1, Fig. S1). We used real environmental variables measured over

137   a real geographic space, so the variety of spatial patterns in our predictor variables should be

138   similar to patterns in variables that determine biological species distributions at this scale, adding

139   realism to our simulation. We used climate variables (which show relatively strong spatial

140   clustering, Table 1) from the E-OBS European Climate Assessment and Dataset EU project

141   (Haylock et al., 2008; van den Besselaar, Haylock, van der Schrier, & Klein Tank, 2011). We

142   calculated the proportion of each grid square covered by different land cover variables (which

143   show less spatial clustering than climate variables, Table 1) from the CORINE Land Cover

144   database (CORINE, 2012). We calculated the average elevation within each grid square by

145   interpolation using ordinary kriging from the ETOPO1 Global Relief Model (Amante & Eakins,

146   2009).

147   Spatial data were prepared using the 'sf', 'sp', 'raster', 'fasterize', 'rgdal', 'gstat', and

148   'tidyverse' packages in R version 3.6 (Bivand, Keitt, & Rowlingson, 2018; Gräler, Pebesma, &

149   Heuvelink, 2016; Hijmans 2018; Pebesma, 2018; R Core Team, 2020; Ross, 2018; Wickham,

150   2017).

151   **2.2   Species presence data**

152   We downloaded observations of species across the island of Ireland for the years 1970 to 2014

153   from the British Bryological Society for bryophytes and from the Irish National Biodiversity

154   Data Centre (NBDC) for moths, butterflies, and birds. We used data for taxa that varied in both

155   their popularity with recorders and the ease with which species can be identified, because we

156   expected this would translate into different spatial sampling biases in the data. The data

157   contained presence-only records of species, with the date and location of the observation, and an

158   anonymized observer identifier. Locations of records were provided as either 1 km^2 or 100 km^2

159   (10 km x 10 km) grid squares, but we used 10 km x 10 km grid squares in all analyses in order to

160   retain the majority of the data. Within each taxonomic group, we grouped records into sampling

161   event checklists, where a sampling event was defined as all records with an identical

162   combination of recording date, location, and observer.

163   **2.3   Spatial sampling patterns in Irish species presence data**

164   For each taxonomic group, we quantified sampling effort in each grid square as the proportion

165   of all records coming from the grid square. We quantified spatial sampling bias by calculating

166   the spatial evenness of sampling effort among locations using Simpson evenness (Magurran &

167   McGill, 2011).

168   **2.4   Data simulation**

169   *2.4.1   Simulating species distributions*

170   We simulated and sampled virtual species distributions using the 'virtualspecies' package

171   (Leroy et al., 2016) in R. For each virtual species, seven environmental variables (Table 1) were

172   randomly chosen to use as drivers of occurrence (only seven of the ten variables shown in Table

173   1 were used for each species so that not all species responded to exactly the same environmental

174   variables). The seven selected environmental variables were centered, scaled, and summarized

175   using principal components analysis with the 'ade4' R package (Dray & Dufour, 2007). The

176   probability of occurrence of each virtual species *i* in each grid square *j* was a logistic function of

177   the first two principal components and their quadratic terms:

178   $logit(p_{ij}) = \alpha_i + \sum_{k=1}^{2}(\beta_{1ki}V_{kj} + \beta_{2ki}V_{kj}^2)$

179   where p_{ij} is the probability that virtual species *i* occurs in grid square *j*, V_{kj} indicates the value of

180   the kth principal component in grid cell *j,* and the α and β terms are the species-specific

181   coefficients defining the response of the virtual species to the environment.   The coefficients for

182   each virtual species (α, β_{1k}, β_{2k}) were chosen to ensure that each virtual species was present in

183   at least eight of the 840 grid squares.

184   *2.4.2   Realized species distributions*

185   A single realized distribution of each virtual species i was created by randomly generating a

186   "presence" (1) or "absence" (0) for each grid square j by drawing a value from a binomial

187   distribution with probability p_{ij}. We simulated a community containing 1268 virtual species (the

188   number of recorded bryophyte species in Ireland).   For comparison, results of a simulation using

189   a small community of 34 species are in Supplementary Article S1).

190   *2.4.3   Sampling realized species with spatial bias*

191   Virtual biological records data were generated by sampling the realized species communities in

192   "sampling events" at different locations to produce spatially explicit species checklists (Fig. S2).

193   Spatial sampling locations were chosen based on spatial sampling patterns from three Irish

194   biological records datasets with different spatial sampling biases (Table 2): birds (low spatial

195   sampling bias), butterflies (median spatial sampling bias), and moths (severe spatial sampling

196   bias). This gave four spatial sampling "templates", including the case of no spatial sampling bias

197   (Fig. 2).

198   To make sampling patterns comparable between datasets with different sample sizes, we

199   calculated a relative sampling weight for each grid square in each empirical dataset (where the

200   most heavily sampled cell had a weight of one) by counting the number of records in each grid

201   square and dividing by the maximum number of records in any grid square (Fig. 2).

202   We created virtual biological records with six different sample sizes, defined as the mean

203   number of records per species (number of records per species = 2, 5, 10, 50, 100, and 200).

204   To generate virtual biological records from the virtual species communities, we randomly

205   selected a grid square, using selection probabilities from one of the four spatial-bias templates.

206   Within each grid square that was selected for sampling, we 1) generated a list of virtual species

207   that were present in the grid square; 2) defined the probability of observing each of the present

208   species based on the species' prevalence in the entire study extent (so that common species had a

209   higher probability of being recorded when present), and 3) drew presence observations with

210   replacement from the list of present species. We continued this sampling process until we had

211   accumulated the desired number of records.

212   **2.5   Species distribution modeling**

213   We tested three different SDM modeling techniques: generalized linear models (GLMs)

214   (Hosmer & Lemeshow, 2000), boosted regression trees (Elith, Leathwick, & Hastie, 2008;

215   Friedman, 2001), and inverse distance-weighted interpolation (Cressie, 1991). These represent

216   distinct types of methods used for SDMs, including linear and machine learning methods, and a

217   spatial interpolation method that does not include information from environmental covariates.

218   For all methods, the modeled quantity was the probability of the focal virtual species being

219   recorded on a checklist. We modeled each species individually as a function of five

220   environmental predictor variables, chosen from the ten possible predictor variables listed in

221   Table 1. Using only five of the ten possible predictor variables simulated a real-world situation in

222   which the factors that influence species distributions are not entirely known. We treated the list

223   of records from each sampling event as a complete record of that sampling event, and treated the

224   absence of species from a sampling event checklist as non-detection data for those species (Fig.

225   S2, van Strien, van Swaay, & Termaat, 2013; Kéry et al., 2010). Thus, we explicitly used a

226   detection/non-detection rather than a presence-only modeling framework. Our approach of

227   inferring non-detections of species at the locations of presence records of other species in the

228   community is similar to the "target-group background" approach of Phillips et al. (2009). Using

229   non-detection data inferred from records of other species ensured that the sampling biases were

230   the same for detections and non-detections, which may reduce the effect of sampling bias

231   (Barbet-Massin et al. 2012, Johnston et al. 2020, Phillips et al. 2009).

232   We modeled 110 randomly selected virtual species from the 1268 virtual species in the large

233   community simulation. The number of virtual species modeled was a compromise between high

234   replication and computation limitations, but testing 110 virtual species should provide enough

235   replication for robust conclusions. We fitted each type of SDM once to each combination of

236   virtual species, sample size, and spatial sampling bias. Thus, the sample size for our study – the

237   number of SDM prediction performance values that we used to assess the effects of spatial

238   sampling bias, sample size, and SDM method - was 110 prediction performance values for each

239   combination of SDM method, sample size, and spatial sampling bias. Replication in our study

240   came not from repeatedly fitting models to different randomly generated sets of presences and

241   absences of the same virtual species, but rather from fitting each model once to data for many

242   different virtual species, all generated using parameters randomly drawn from the same

243   distributions. However, the same 110 virtual species were used for each combination of SDM

244   method, spatial sampling bias, and sample size, ensuring that all comparisons were based on the

245   same underlying task (i.e. modelling the same true species distributions).

246   Models were trained and evaluated using five-fold spatial block cross-validation (Roberts et al.,

247   2017) that partitioned the study extent into spatial blocks of 100 km x 100 km and allocated each

248   block to one of five cross-validation partitions. We only attempted to fit models if there were

249   more than five presence records in the training data. Prediction performance of models was

250   evaluated using the true virtual species presence or absence in each grid cell not included in the

251   spatial extent of the training partitions (Fig. 3). Thus, evaluation data was spatially even and the

252   number of evaluation points stayed constant even as the sample size and spatial bias of training

253   data changed (Fig. 3). Prediction performance was evaluated using three metrics. The area

254   under the receiver operating characteristic curve (AUC) (Hosmer & Lemeshow, 2000) evaluated

255   models' ability to accurately rank locations where species were present or absent using the

256   continuous SDM predictions, Cohen's Kappa (Cohen 1960), calculated using the threshold that

257   maximised Kappa, evaluated models ability to transform continuous SDM predictions into

258   binary maps of presence and absence, and root mean squared error (RMSE) evaluated model

259   calibration.

260   To train GLM SDMs, we used logistic regression ('glm' function) with a binomial error

261   distribution and logit link. Quadratic terms were fitted, but we did not fit interactions between

262   variables. We controlled overfitting by limiting the number of terms in GLMs such that there

263   were at least 10 detections or non-detections (whichever was smaller) in the training data for

264   each non-intercept term in the model. If a quadratic term was included in a model, we also

265   included the 1st degree term. For generating predictions, we used the model that gave the lowest

266   AIC based on the training data.

267   Boosted regression trees were trained using 'gbm.step' in the 'dismo' package (Greenwell,

268   Boehmke, & Cunningham, 2018; Hijmans, Phillips, Leathwick, & Elith, 2017). Unlike GLMs,

269   boosted regression trees do not require the modeler to specify interactions between variables,

270   because the trees will discover and model interactions if they are present. The tree complexity

271 specified by the modeler controls the maximum interaction order that the models are permitted to

272 fit, and therefore can be used to prevent overfitting. We tested tree complexities of two and five,

273 to build models that allowed interactions between up to two and up to five variables,

274 respectively. We used learning rates small enough to grow at least 1000 trees (following Elith,

275 Leathwick, and Hastie 2008), but large enough to keep models below an upper limit of 30,000

276 trees because of computation time limitations. We used gbm.step to determine the optimal

277 number of trees for each model, based on monitoring the change in 10-fold cross-validated error

278 rate as trees were added to the model (Hijmans, Phillips, Leathwick, & Elith, 2017). Limiting

279 models to a maximum of 30,000 trees did not affect our results. Details of the procedure are in

280 Supplementary Article S1 and in our R code (see Data and code accessibility statement).

281   Inverse distance-weighted interpolation was implemented using 'gstat' (Gräler et al., 2016;

282 Pebesma, 2004). We tuned parameters of the inverse distance-weighted interpolation model

283 based on prediction error (details in Supplementary Article S1).

284   After models were fitted, we looked for evidence of overfitting and assessed whether the

285 number of presence records of the focal species in the test dataset affected prediction

286 performance metrics (details in Supplementary Article S1). All analyses used R version 3.6.0 (R

287 Core Team, 2020).

288 **2.6   Analyzing effects of sampling bias and sample size**

289   We modeled the predictive performance (AUC, maximum Cohen's Kappa, and RMSE) of

290 SDMs as a function of spatial sampling bias, sample size (average number of observations per

291 species), and SDM method. Modelling was done using boosted regression trees ('gbm.step' in

292 the 'dismo' package) (Greenwell et al., 2018; Hijmans et al., 2017). To assess whether species

293   prevalence (the commonness or rarity of a species in the study extent) affected our evaluations of

294   model performance, we graphed prediction performance as a function of species prevalence for

295   all models (Fig. S3). We included species prevalence in the boosted regression tree models of

296   RMSE, but not in our analysis of AUC or Cohen's Kappa, because only RMSE showed a strong

297   trend with species prevalence (Fig. S3). Variable importance was assessed based on the

298   reduction in squared error attributed to each variable in boosted regression tree models

299   (Friedman, 2001). We also assessed the effect of spatial sampling bias and sample size of

300   training data on the number of species for which models could be fitted within the computational

301   time and memory constraints of this study (Supplementary Article S1).

302   **3   RESULTS**

303   Simulated species showed a variety of plausible distribution patterns (Fig. 4) and prevalences

304   (Fig. S4), including species with north/south distribution gradients and distributions that

305   followed geographic features such as the coastline (Fig. 4).

306   Sample size (the mean number of observations per species) and choice of SDM method were

307   the most important variables for explaining variations in prediction performance of SDMs (Table

308   3). Spatial sampling bias was the least important variable for explaining variation in prediction

309   performance for all three performance metrics (Table 3).   Simpson evenness values for spatial

310   sampling evenness of the template datasets are in Table 2.

311   **3.1   Number of species successfully modeled**

312   The number of species for which models fitted successfully increased as sample size increased

313   and spatial bias decreased (Fig. 5). For GLMs and inverse distance-weighted interpolation,

314   model fitting was largely successful when datasets had more than 100 records per species, except

315   when spatial bias was severe (Fig. 5). Boosted regression trees failed to fit models for some

316   species even with relatively large amounts of data (e.g. an average 200 records per species), and

317   models fit less frequently when data had median or severe spatial biases (Fig. 5). The effect of

318   spatial bias on the number of species for which models fitted was small, but was slightly greater

319   for boosted regression trees than for other SDM modelling methods (Fig. 5).

320   **3.2   Predictive performance of SDMs**

321   The amount of spatial bias in training data was less important than sample size and choice of

322   SDM method in predicting the performance of SDMs (Table 3, Table S1). AUC for predictive

323   SDMs increased with the average number of records per species and with decreasing spatial bias

324   in the training data when using all SDM methods (Fig. 6, Fig. 7). Root mean squared error

325   (RMSE) was largely unaffected by spatial sampling bias (Fig. 8, Table 3). Cohen's Kappa

326   differed between SDM methods, but generally increased with sample size, and with decreasing

327   spatial bias in the training data (Fig. S5, Fig. S6). However, Cohen's Kappa was low for all

328   methods and combinations of sample size and spatial bias when evaluated using spatial block

329   cross-validation (Fig. S6). The low Kappa scores indicated that our models were unable to

330   generalize beyond the training data when the task was to produce binary maps, despite the fact

331   that the models retained the ability to generalize when the task was to rank sites according to the

332   continuous SDM outputs (Supplementary Article S1, Fig. S7). Species prevalence (the number

333   of grid squares occupied by a species) had a negligible effect on the average value of AUC,

334   though it did affect the variability of AUC (Fig. S3). Species prevalence strongly affected the

335   expected value of RMSE, with RMSE increasing with species prevalence (Table 3, Fig. S3).

336   *3.2.1   Effect of sample size*

337   Sample size (average number of records per species) was the most important variable for

338   predicting species distribution model prediction performance in terms of AUC (Table 3). AUC

339   improved with increasing average number of records per species for all SDM methods, and the

340   improvement in AUC decelerated as the number of records per species increased (Fig. 6, Fig. 9).

341   Kappa improved with increasing average number of records per species for two of the three

342   SDM methods, and the improvement in Kappa decelerated as the number of records per species

343   increased (Fig. S5).

344   *3.2.2   Effect of spatial bias*

345   Higher levels of spatial sampling bias generally reduced AUC and Kappa, but the size of this

346   effect was small for the low level of bias (Fig. 6). SDMs built with GLMs showed the biggest

347   difference in prediction performance between models trained with unbiased data and models

348   trained with data showing median spatial bias (reduction in expected AUC of 0.037 when using

349   an average of 200 records per species, Fig. 6). Other SDM methods showed less difference in

350   AUC between models trained with unbiased data and models trained with data containing

351   median spatial bias (decrease in expected AUC of 0.033 for boosted regression trees and 0.030

352   for inverse distance-weighted interpolation when using an average of 200 records per species).

353   The AUC for inverse distance-weighted interpolation models trained with unbiased data was

354   generally higher than the AUC for GLMs and boosted regression trees trained with severely

355   biased data, but lower than the AUC for GLMs and boosted regression trees trained with data

356   with median spatial bias for any given sample size (Fig. 6, Fig. 7).

4   DISCUSSION

357

358   Both sample size (the average number of observations per species) and choice of modelling

359   method were more important than the spatial bias of training data for determining model

360   prediction performance. This is in line with the results of Thibaud et al. (2014), and the

361   importance of sample size was in line with studies summarized in Meynard, Leroy & Kaplan

362   (2019). Thibaud et al. (2014) simulated spatial sampling bias by defining sampling probability

363   as a linear function of distance from the nearest road. In contrast, our study used observed

364   spatial sampling patterns from real biological records datasets. Our results therefore provide a

365   more direct confirmation that spatial biases of the type and intensity found in real datasets are not

366   as important as other factors in determining SDM prediction performance. We did not measure

367   how different our biased sampling was from Thibaud et al.'s (2014) biased sampling, but we

368   suspect biases in real data are more severe than the bias tested by Thibaud et al. (2014). More

369   attempts to quantify sampling biases in real data would be valuable for informing how bias is

370   incorporated into future simulation studies.

371   While spatial bias was not the most important factor determining SDM prediction performance,

372   spatial sampling bias did affect model prediction performance when spatial bias was relatively

373   strong. The limited effect of spatial bias on SDMs that we observed is similar to other findings

374   that have shown spatial sampling bias to have a small effect on model performance (Thibaud et

375   al., 2014; Warton et al., 2013) or to affect only some SDM methods (Barbet-Massin et al., 2012).

376   Our study used spatial biases and the spatially explicit environmental data representative of data

377   likely to be used in SDMs using biological records in Ireland. Our conclusions therefore apply

378   most directly to applications of SDMs using Irish biological records, and may not be

379   generalizable to other geographic locations, or for species within Ireland that do not respond to

380   the environmental predictor variables used in this study. Our simulations could be "scaled up" to

381   use environmental variables from, and create virtual species over, a larger spatial extent, which

382   would provide insight about whether spatial sampling bias remains relatively un-important at

383   larger (e.g. continental) scales. However, our results strengthen a growing body of literature that

384   suggests that spatial sampling bias is rarely the most important issue in determining SDM

385   prediction performance. In particular, the choice of modelling method may often have more

386   impact on SDM prediction performance than a variety of other factors (Barbet-Massin et al.,

387   2012; Fernandes, Scherrer, & Guisan, 2018).

388   Training data with low spatial sampling bias produced species distribution models that

389   performed nearly as well as models trained with unbiased data, when the task was ranking the

390   probability of a species being detected at sites. Prediction performance was poor when models

391   were trained with small sample sizes, regardless of the spatial bias in training data. Similarly,

392   model performance increased quickly with sample size when sample size was small, even when

393   the data had severe spatial bias. This suggests that, for taxonomic groups with relatively few

394   records per species, the usefulness of the data for predictive SDMs can be improved by

395   increasing sample size, even if additional data collection is spatially biased. In contrast, for

396   taxonomic groups for which biological records datasets already have a high average number of

397   records per species (e.g. birds and butterflies which both have an average of over 2000 records

398   per species in Ireland) further improvements in SDM prediction performance will likely require

399   increasing the spatial evenness of data (Fig. 9).

400   While our SDMs were able to correctly rank sites in terms of the probability of a focal species

401   being detected (measured using AUC), the SDMs were essentially unable to convert the

402   continuous SDM outputs into binary presence/absence maps (measured using Cohen's Kappa)

403   for locations outside the training data (Fig. S6). Converting continuous SDM outputs to binary

404   classifications is fraught with difficulties, and the criteria that should be used for selecting the

405   threshold for conversion depend on the purpose of the SDMs (Guillera-Arroita et al., 2015, and

406   references therein). Any conversion of a continuous SDM output to a binary classification

407   necessarily results in a loss of information, and Guillera-Arroita et al. (2015) suggest that for

408   most applications of SDMs, binary conversions are not necessary. It was outside the scope of

409   this study to explore why SDMs were better at ranking than at producing binary classifications

410   for sites outside the training data. The main finding of our study, that spatial sampling bias was

411   less important than other factors in determining model performance, was consistent regardless of

412   whether prediction performance was evaluated in terms of the ability to rank (AUC) or classify

413   (Kappa) sites (Table 3).

414   Presence-only data are common in biological records, but there are very few true "presence-

415   only" SDM modeling methods. Most common SDM methods require information on presences

416   *and* something else (e.g. non-detections, or background- , quadrature-, or pseudo-absences

417   points) (Phillipes et al., 2009; Elith, Phillips, Hastie, Dudik, Chee & Yates, 2011; Renner et al.,

418   2015). We inferred non-detections from detections of other species (e.g. Phillips et al., 2009;

419   van Strien, Termaat, Groenendijk, Mensing, & Kéry, 2010; Isaac, van Strien, August, de Zeeuw,

420   & Roy, 2014; Johnston et al. 2020). Other options include generating pseudo-absences randomly

421   or according to a variety of rules (Barbet-Massin et al. 2012). Phillips et al. (2009) proposed a

422   "target group" approach that selects background points using detections of other species. The

423   target group approach is essentially identical to ours. Our results – particularly the finding that

424   spatial sampling bias was relatively un-important relative to other factors – applies most directly

425   to SDM approaches that infer non-detections from presence records of other species. When

426   background or pseudo-absence points are generated randomly or using an ad-hoc rule, we expect

427   that SDMs will be more strongly impacted by spatial bias in the presence-only data than were

428   SDMs in our study (Phillips et al. 2009).

429   The objective of our SDMs was to fill in gaps in species distribution knowledge within the

430   spatial and environmental conditions of the island of Ireland, an area of about 84,000 km^2. Our

431   results may not generalize to larger spatial scales or to cases in which the goal of SDMs is

432   uncovering species' entire fundamental environmental niche or determining the environmental

433   factors most strongly influencing distributions. The spatial scope of our SDMs is sensible both

434   from an ecological and applied standpoint, because the island of Ireland is a geographically

435   delimited ecological unit, and because decision making about species conservation and

436   management often happens within political units (e.g. nations, states, or counties) that cover only

437   a portion of species' spatial and environmental distributions.

438   GLMs had the best prediction performance of the SDM methods we tested, even though they

439   were more affected by spatial bias than were other methods. The high performance of GLMs

440   relative to other modelling methods in this study agrees with the simulation results of Thibaud et

441   al. (2014) and Fernandes et al. (2018). However, as in both those studies, we generated virtual

442   species distributions according to a linear model, so it is possible that the good performance of

443   GLMs is due to the model having the same functional form as the "true" species responses. In

444   real applications, it is unlikely that the functional form of the model will exactly match the form

445   of the true species responses. Indeed, the species distribution modelling literature has many

446   examples of different modelling methods performing best in different studies, suggesting that no

447   modelling method consistently outperforms others (Bahn & McGill, 2007; Breiner, Nobis,

448   Bergamini, & Guisan, 2018; Cutler et al., 2007; Elith et al., 2006; Elith & Graham 2009).

449   Boosted regression trees' prediction performance was slightly less affected by spatial bias than

450   GLMs', and prediction performance of both methods was similar when trained with large,

451   spatially biased datasets. But boosted regression trees failed to fit models more often than did

452   GLMs, especially when sample sizes were smaller, which may make them inferior to other

453   modelling methods for small datasets, at least within the computational resource limits we faced.

454   We cannot rule out the possibility that the performance of boosted regression trees would

455   improve if they were trained with a smaller learning rate and permitted to grow more than 30,000

456   trees. However, most users of SDMs will face some computational resource limitations.

457   In this study, we introduced spatial bias specifically into the training data and tested model

458   performance using spatially even evaluation data. However, spatial bias can also occur in

459   evaluation data and may affect the reliability of model evaluations (Fink et al., 2010). When

460   using real biological records datasets, it is likely that both model training and evaluation will use

461   spatially biased data, making it difficult to dis-entangle whether observed effects of spatially

462   biased data on prediction performance are due to the influence of biased data in the model

463   training step or in the model evaluation step. We evaluated models on spatially even data (which

464   is easy using simulated data but would be more difficult or impossible when using real data), so

465   the observed effects of spatially biased data on prediction performance in our study can be

466   attributed to the effect of biased data on model training. All of the SDM methods we used

467   involve some kind of model evaluation as part of the model training process, either inherent in

468   the model fitting or introduced by our implementation. For example, with our GLMs we

469   introduced a model evaluation step to select predictor variables. The final GLM models were

470   therefore based on variables that had been selected by evaluation on spatially biased data. For

471   both GLMs and inverse distance-weighted interpolation, it is possible that using unbiased data in

472   the evaluations during model selection would have led to different final models. Therefore, the

473   observed effect of the spatial bias in this study could be due to how biased data affects the actual

474   fitting of each individual model, or to how the biased data affects the evaluation step used to

475   select which fitted model to use for predictions. Tree-based methods, including boosted

476   regression trees, select which values of predictor variables to split at and/or which predictor

477   variables to use at each node based on how much those splits improve some measure of

478   performance on the training data (Elith et al., 2008; Hastie et al., 2009). Thus, evaluation on

479   potentially spatially biased training data is inherent in fitting tree models.

480   Fink et al. (2010) provided a method for correcting spatial bias in evaluation data to reduce the

481   effect of spatial bias on model evaluation, but they did not explicitly address spatially biased data

482   in model training. Our results showed that spatially biased data can impact model training (at

483   least when the spatial bias is relatively strong). Investigating the effect of spatially biased data on

484   the evaluation that takes place as part of model training (e.g. during variable selection or

485   parameter tuning) may be a worthwhile path for future research. It may be possible to use a

486   method like that proposed by Fink et al. (2010) to correct spatial bias during the evaluation that

487   takes place within the model training process.

488

489   **5   CONCLUSION**

490   We found that spatial sampling bias in training data affected species distribution model

491   prediction performance when the spatial bias was relatively strong, but that sample size and the

492   choice of modelling method were more important than spatial bias in determining model

493   prediction performance. This study adds to a body of literature suggesting that prediction

494   performance of species distribution models is less affected by spatial sampling bias in training

495   data than by other factors including modelling method and sample size. We suggest that, when

496   biological records datasets are relatively small, model prediction performance can best be

497   improved by increasing the number of records, even if additional data are sampled with spatial

498   bias. Attempts to reduce spatial bias in data through data filtering should be cautious about the

499   resulting decrease in sample size, which could cancel any gains in prediction performance from

500   reducing spatial sampling bias. Converting continuous SDM outputs to binary maps should be

501   done with caution, and prediction performance tests should use spatial block cross-validation or

502   test on independent data.

503   **6   ACKNOWLEDGMENTS**

504   We thank Tomás Murray and the Irish National Biodiversity Data Centre (NBDC) for

505   providing and answering questions about the biological records data, and we thank the many

506   citizen scientists who collected and contributed their data to the NBDC. This work used the

507   ResearchIT Sonic cluster funded by UCD IT Services and the Research Office. This research

508   used CORINE data made available with funding by the European Union.

509

510

511

512   **REFERENCES**

513   Amano, T., & Sutherland, W. J. (2013). Four barriers to the global understanding of biodiversity

514   conservation: Wealth, language, geographical location and security. *Proceedings of the Royal*

515   *Society B: Biological Sciences,* 280, 20122649.

516   Amano, T., Lamming, J. D. L., & Sutherland, W. J. (2016). Spatial gaps in global biodiversity

517   information and the role of citizen science. *BioScience,* 66, 393–400.

518   Amante, C., & Eakins, B. W. (2009). ETOPO1 1 arc-minute global relief model: Procedures,

519   data sources and analysis. NOAA Technical Memorandum NESDIS NGDC-24. National

520   Geophysical Data Center, NOAA. doi: 10.7289/V5C8276M [accessed 8 May 2019].

521

522   Bahn, V., & McGill, B. J. (2007). Can niche-based distribution models outperform spatial

523   interpolation? *Global Ecology and Biogeography,* 16, 733–742.

524   Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences

525   for species distribution models: How, where and how many? *Methods in Ecology and*

526   *Evolution,* 3, 327–338.

527   Bivand, R., Keitt, T., & Rowlingson, B. (2018). *rgdal: Bindings for the 'geospatial' data*

528   *abstraction library*. R package versions 1.3-9 and 1.4-4.

529   Boakes, E. H., McGowan, P. J. K., Fuller, R. A., Chang-qing, D., Clark, N. E., O'Connor, K., &

530   Mace, G. M. (2010). Distorted views of biodiversity: Spatial and temporal bias in species

531   occurrence data. *PLoS Biology,* 8, e1000385.

532   Breiner, F. T., Nobis, M. P., Bergamini, A., & Guisan, A. (2018). Optimizing ensembles of small

533   models for predicting the distribution of species with few occurrences. *Methods in Ecology*

534   *and Evolution,* 9, 802–808.

535   Carvell, C., Isaac, N. J. B., Jitlal, M., Peyton, J., Powney, G. D., Roy, D. B., ... Roy, H. E.

536   (2016). *Design and testing of a national pollinator and pollination monitoring framework.*

537   Final summary report to the Department for Environment, Food and Rural Affairs (Defra),

538   Scottish Government; Welsh Government: Project WC1101.

539   Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological*

540   *Measurement,* 20, 37-46.

541   CORINE land cover database. (2012). Version 18. © European Union, Copernicus Land

542   Monitoring Service 2016, European Environment Agency (EEA). Retreived from

543   https://www.eea.europa.eu/ds_resolveuid/ecb838dabf4849838ba5f3dc81ca6b0e [8 Aug 2016].

544   Cressie, N. A. C. (1991). *Statistics for spatial data.* New York: John Wiley & Sons, Inc.

545   Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J.

546   (2007). Random forests for classification in ecology. *Ecology,* 88, 2783–2792.

547   Deka, M., & Morshed, N. (2018). Mapping disease transmission risk of Nipah Virus in South

548   and Southeast Asia. *Tropical Medicine and Infectious Disease,* 3, 57.

549   Dray, S., & Dufour, A. (2007). The ade4 package: Implementing the duality diagram for

550   ecologists. *Journal of Statistical Software,* 22, 1–20.

551   Dyer, R. J., Gillings, S., Pywell, R. F., Fox, R., Roy, D. B., & Oliver, T. H. (2017). Developing a

552   biodiversity-based indicator for large-scale environmental assessment: A case study of

553   proposed shale gas extraction sites in Britain. *Journal of Applied Ecology,* 54, 872–882.

554   Edwards, T. C., Cutler, D. R., Zimmermann, N. E., Geiser, L., & Moisen, G. G. (2006). Effects

555   of sample survey design on the accuracy of classification tree models in species distribution

556   models. *Ecological Modelling,* 199, 132–141.

557   Elith, J., & Graham, C. H. (2009). Do they? How do they? Why do they differ? On finding

558   reasons for differing performances of species distribution models. *Ecography,* 32, 66–77.

559   Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., ... Zimmermann, N.

560   E. (2006). Novel methods improve prediction of species' distributions from occurrence data.

561   *Ecography,* 29, 129–151.

562   Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees.

563   *Journal of Animal Ecology,* 77, 802–813.

564   Elith, J., Phillips, S. J., Hastie, T., Dudik, M., Chee, Y. E., Yates, C. J. (2011). A statistical

565   explanation of MaxEnt for ecologists. *Diversity and Distributions,* 17, 43-57.

566   Fan, W., Davidson, I., Zadrozny, B., & Yu, P. S. (2005). *An improved categorization of*

567   *classifier's sensitivity on sample selection bias*. In Fifth IEEE International Conference on

568   Data Mining (ICDM'05), Houston, TX: IEEE.

569   Fernandes, R. F., Scherrer, D., & Guisan, A. (2018). How much should one sample to accurately

570   predict the distribution of species assemblages? A virtual community approach. *Ecological*

571   *Informatics,* 48, 125–134.

572   Fink, D., Hochachka, W. M., Zuckerberg, B., Winkler, D. W., Shaby, B., Munson, M. A., ...

573   Kelling, S. (2010). Spatiotemporal exploratory models for broad-scale survey data. *Ecological*

574   *Applications,* 20, 2131–2147.

575   Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The*

576   *Annals of Statistics*, 29, 1189–1232.

577   Gelman, A., & Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models

578   (Analytical Methods for Social Research). Cambridge: Cambridge University Press.

579   Gräler, B., Pebesma, E., & Heuvelink, G. (2016). Spatio-temporal interpolation using gstat. *The*

580   *R Journal,* 8, 204–218.

581   Greenwell, B., Boehmke, B., Cunningham, J., & GBM Developers. (2018). g*bm: Generalized*

582   *boosted regression models*. R package version 2.1.4.

583   Guillera-Arriota, G., Lahoz-Monfort, J. J. Elith, J. Gordon, A. Kujala, H., Lentini, P. E.,

584   McCarthy, M. A., Tingley, R., Wintle, B. A. (2015). Is my species distribution model fit for

585   purpose? Matching data and models to applications. *Global Ecology and Biogeography,* 24,

586   276-292.

587   Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data*

588   *mining, inference and prediction* (2nd ed.). New York: Springer.

589   Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., & New, M. (2008).

590   A European daily high-resolution gridded data set of surface temperature and precipitation for

591   1950-2006. *Journal of Geophysical Research,* 113, D20119.

592   Hijmans, R. J. (2018). *raster: Geographic data analysis and modeling*. R package versions 2.8-4

593   and 2.9-23.

594   Hijmans, R. J., Phillips, S., Leathwick, J., & Elith, J. (2017). *dismo: Species distribution*

595   *modeling*. R package version 1.1-4.

596   Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression (*2nd ed). New York:

597   Wiley.

598   Isaac, N. J. B., and Pocock, M. J. O. (2015). Bias and information in biological records.

599   *Biological Journal of the Linnean Society,* 115, 522–531.

600   Isaac, N. J. B., van Strien, A. J., August, T. A., de Zeeuw, M. P., & Roy, D. B. (2014). Statistics

601   for citizen science: Extracting signals of change from noisy ecological data. *Methods in*

602   *Ecology and Evolution,* 5, 1052–1060.

603   Johnston, A., Moran, N., Musgrove, A., Fink, D., & Baillie, S. R. (2020). Estimating species

604   distributions from spatially biased citizen science data. *Ecological Modelling*, 422, 108927.

605   https://doi.org/10.1016/j.ecolmodel.2019.108927

606   Kéry, M., Royle, A., Schmid, H., Schaub, M., Volet, B., Häfliger, G., & Zbinden, N. (2010).

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

608   opportunistic observations. *Conservation Biology*, 24, 1388-1397.

609   Leroy, B., Meynard, C. N., Bellard, C., & Courchamp, F. (2016). virtualspecies, an R package to

610   generate virtual species distributions. *Ecography,* 39, 599–607.

611

612   Magurran, A. E., & McGill, B. J. (Eds.). (2011). *Biological diversity: Frontiers in measurement*

613   *and assessment*. Oxford: Oxford University Press.

614   Meynard, C. N., Leroy, B., & Kaplan, D. M. (2019). Testing methods in species distribution

615   modelling using virtual species: what have we learnt and what are we missing?. *Ecography,*

616   42, 2021-2036.

617   Nakagawa, S. (2015). Missing data: mechanisms, methods, and messages. In G. A. Fox, S.

618   Negrette-Yankelevich, & V. J. Sosa (Eds.), *Ecological statistics: Contemporary theory and*

619   *application* (First Ed., pp. 81–105). Oxford: Oxford University Press.

620   Newson, S. E., Evans, H. E., Gillings, S., Jarrett, D., Raynor, R., & Wilson, M. W. (2017).

621   Large-scale citizen science improves assessment of risk posed by wind farms to bats in

622   southern Scotland. *Biological Conservation,* 215, 61–71.

623 Oliveira, U., Paglia, A. P., Brescovit, A. D., de Carvalho, C. J. B., Silva, D. P., Rezende, D. T., …

624   Santos, A. J. (2016). The strong influence of collection bias on biodiversity knowledge

625   shortfalls of Brazilian terrestrial biodiversity. *Diversity and Distributions,* 22, 1232–1244.

626 Pebesma, E. (2018). Simple features for R: Standardized support for spatial vector data. *The R*

627   *Journal,* 10, 439-446.

628 Pebesma, E. J. (2004). Multivariable geostatistics in S: The gstat package. *Computers &*

629   *Geosciences,* 30, 683–691.

630 Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S.

631   (2009). Sample selection bias and presence-only distribution models: Implications for

632   background and pseudo-absence data. *Ecological Applications,* 19, 181–197.

633 R Core Team (2020). *R: A language and environment for statistical computing*. Vienna, Austria:

634   R Foundation for Statistical Computing.

635 Redding, D. W., Atkinson, P. M., Cunningham, A. A., Lo Iacono, G., Moses, L. M., Wood, J. L.

636   N., & Jones, K. E. (2019). Impacts of environmental and socio-economic factors on

637   emergence and epidemic potential of Ebola in Africa. *Nature Communications*, 10, 4531.

638 Renner, I. W., Elith, J., Baddeley, A., Fithian, W. Hastie, T., Phillips, S. J., Popovic, G., &

639   Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology*

640   *and Evolution,* 6, 366-379.

641 Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., … Dormann, C.

642   F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or

643   phylogenetic structure. *Ecography,* 40, 913–929.

644 Ross, N. (2018). *fasterize: Fast polygon to raster conversion*. R package version 1.0.0.

645   Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J. B., & O'Hara, R. B. (2020). Is more

646   data always better? A simulation study of benefits and limitations of integrated distribution

647   models. *Ecography,* 43, 1-10.

648   Stolar, J., & Nielsen, S. E. (2015). Accounting for spatially biased sampling effort in presence-

649   only species distribution modelling. *Diversity and Distributions,* 21, 595–608.

650   Thibaud, E., Petitpierre, B., Broennimann, O., Davison, A. C., & Guisan, A. (2014). Measuring

651   the relative effect of factors affecting species distribution model predictions. *Methods in*

652   *Ecology and Evolution,* 5, 947–955.

653   Uzarski, D. G., Brady, V. J., Cooper, M. J., Wilcox, D. A., Albert, D. A., Axler, R. P., ...

654   Schneider, J. P. (2017). Standardized measures of coastal wetland condition: Implementation

655   at a Laurentian Great Lakes basin-wide scale. *Wetlands,* 37, 15-32.

656   van den Besselaar, E. J. M., Haylock, M. R., van der Schrier, G., & Klein Tank, A. M. G. (2011).

657   A European daily high-resolution observational gridded data set of sea level pressure. *Journal*

658   *of Geophysical Research Atmospheres*, 116, D11110.

659   van Strien, A. J., van Swaay, C. A. M. & Termaat, T. (2013). Opportunistic citizen science data

660   of animal species produce reliable estimates of distribution trends if analysed with occupancy

661   models. *Journal of Applied Ecology,* 50, 1450-1458.

662   van Strien, A. J., Termaat, T., Groenendijk, D., Mensing, V., & Kéry, M. (2010). Site-occupancy

663   models may offer new opportunities for dragonfly monitoring based on daily species lists.

664   *Basic and Applied Ecology,* 11, 495-503.

665   Warton, D. I., Renner, I. W., & Ramp, D. (2013). Model-based control of observer bias for the

666   analysis of presence-only data in ecology. *PLoS ONE,* 8, e79168.

667   Wickham, H. (2017). t*idyverse: Easily install and load the 'tidyverse'*. R package version 1.2.1.

668   Wooldridge, J. M. (2009). *Introductory econometrics: A modern approach* (4th ed.). Mason, OH:

669   South-Western.

670   Zurell, D., Berger, U., Cabral, J. S., Jeltsch, F., Meynard, C. N., Münkemüller, T., ... Grimm, V.

671   (2010). The virtual ecologist approach: Simulating data and observers. *Oikos,* 119, 622–635.

Figure 1

Simulation and analysis process.

Step 1 – define spatial sampling biases: we used the locations of biological records of three different taxa in Ireland to produce rasters of the observed probability of sampling from each 10 x 10 km grid cell in Ireland. *Step 2 – generate virtual species true distributions:* the probability of virtual species occurring in each grid cell was defined as a response to real environmental variables. True presence/absence distributions were generated probabilistically. *Step 3 – sample virtual species with spatial biases:* detection-only observations of species were drawn from sampling locations selected according to the spatial sampling bias rasters generated in Step 1, to create a dataset of virtual detection-only biological records. *Step 4 – Infer non-detections using detections of other species:* non-detections of each species were inferred at locations where other species had been recorded, transforming the detection-only virtual biological records into detection/non-detection data. *Step 5 – train SDMs:* SDMs were trained with the virtual detection/non-detection biological records as the response variable and the real environmental variables as predictors. *Step 6 – test SDMs:* SDM prediction performance was assessed using three metrics, which capture different aspects of models prediction performance. Steps 5 and 6 (blue bounding box) were conducted using a spatial block cross-validation framework illustrated in Fig. 3.

Step 1
Define spatial sampling biases

Real biological records data

Four sampling bias rasters (Fig. 2) defining the probability of sampling from each grid cell

Step 2
Generate virtual species true distributions

Real environmental data

Generate virtual species distributions according to a linear model. The probability of the species occurring is a function of environmental predictors:

$$p = B_0 + B_1 X_1 + \ldots B_k X_k$$

True virtual species distributions (Fig. 4)

Step 3
Sample virtual species with spatial biases

Sample presence-only records of virtual species from locations defined by the sampling bias rasters

Presence-only virtual biological records

Step 4
Infer non-detections using detections of other species

Detection/non-detection virtual biological records

Step 5
Train SDMs

Step 6
Test SDMs

SDMs

Continuous predictions of relative probability of detecting the species

Tests how well predicted distributions match true distributions

AUC
Max Kappa
RMSE

Figure 2

Spatial sampling patterns from Irish biological records.

Spatial sampling patterns from Irish biological records were used as templates to create virtual species records data with varying amounts of spatial bias. Darker shades indicate higher relative probability of sampling from a grid square compared to other grid squares *within the same template*; overall sampling effort is the same for each panel (A) through (D). The most heavily sampled grid square in each spatial bias template has a relative recording effort of one, while a grid square with half as many records as the most heavily sampled square has a relative recording effort of 0.5. Spatial sampling patterns derived from datasets for different taxonomic groups were: (A) no bias (even probability of sampling from every grid square), (B) low bias (based on bird data), (C) median bias (based on butterflies), and (D) severe bias (based on moths).

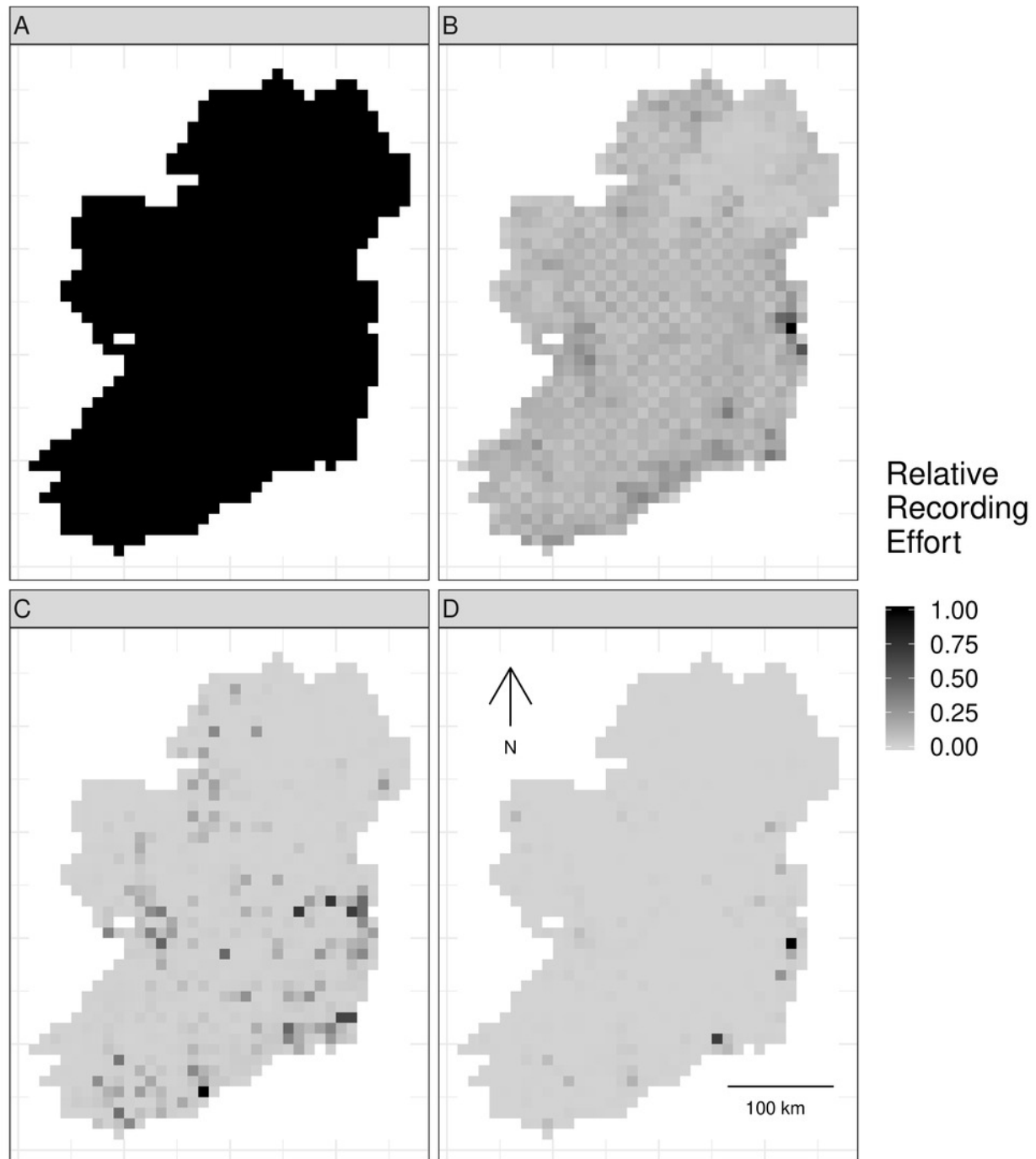



Figure 3

Species distribution model training and testing process for a single cross-validation fold.

The true virtual species distribution (A, presences shown in dark green, absences in light grey) was sampled to produce virtual biological records with a range of sample sizes and spatial biases, including no bias (B) and median bias (C). Orange points in (B) and (C) show checklists on which the species was recorded, black points show checklists on which the species was not recorded (i.e. non-detection points). Species distribution models were fit using five-fold spatial block cross validation, in which data from about 80% of the spatial area was used to train models (light grey background in B and C). Data from the remaining spatial areas (dark grey background in B and C) was set aside for model evaluation. Model evaluation tested the ability of species distribution models to predict the true presence (orange dots) or absence (black dots) of the species in each grid cell within the evaluation areas (D). Model evaluation therefore used spatially even data with the same number of evaluation points (D) regardless of the sample size and spatial bias of training data (B and C).

True virtual
species
distribution

Training data

Evaluation data

A

B

C

D

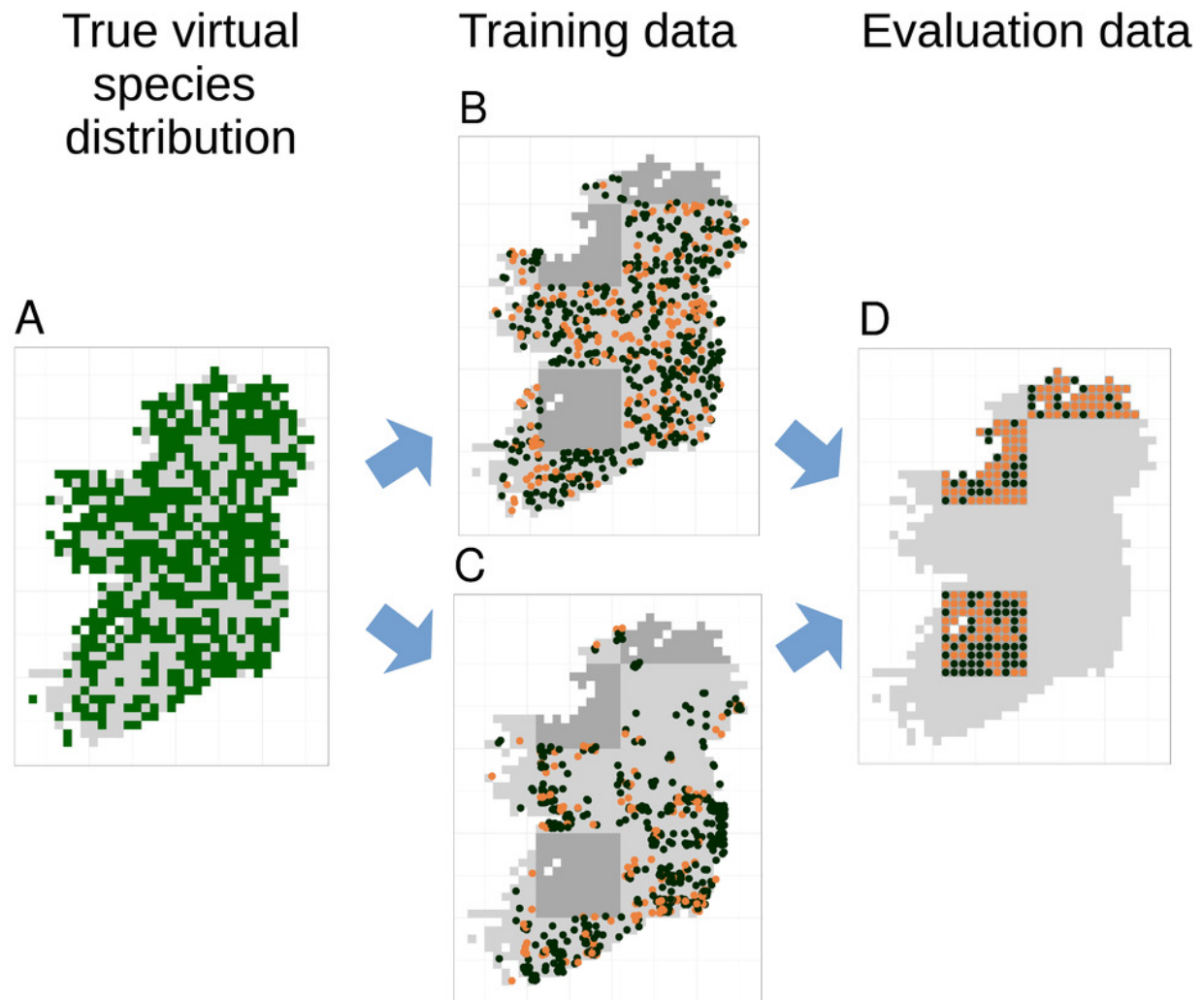
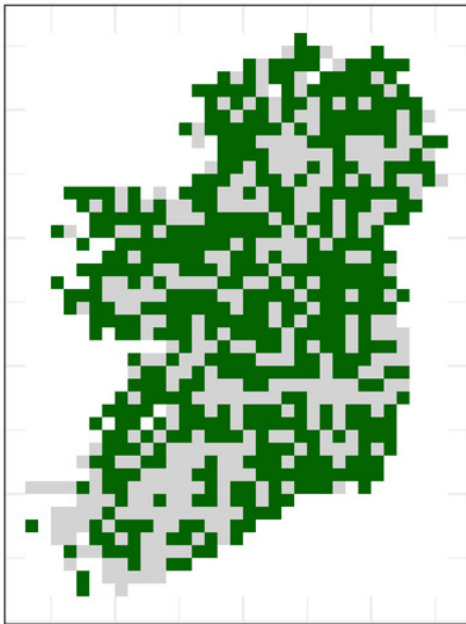
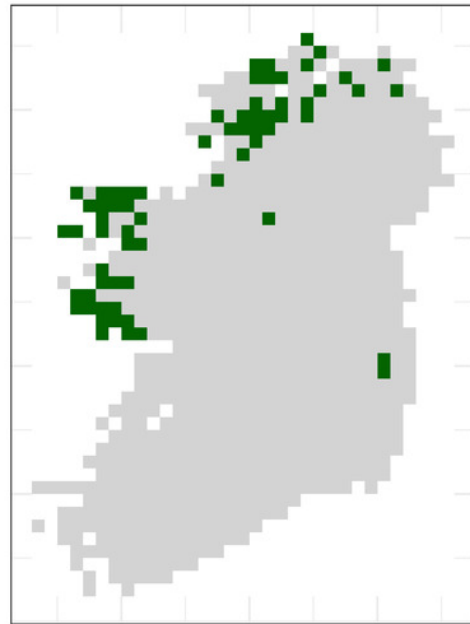
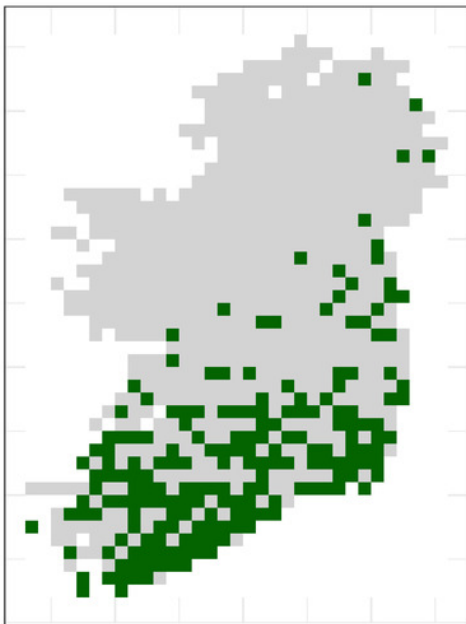



Figure 4

The true distributions of four example simulated species.

Simulated species showed a range of plausible distributions with a range of prevalences, including (A) common widespread species, (B) rare species mostly limited to north-western coastal sites, (C) species with a north/south gradient in occurrence, and (D) common species that are absent from southern sites.

A

B

C

D

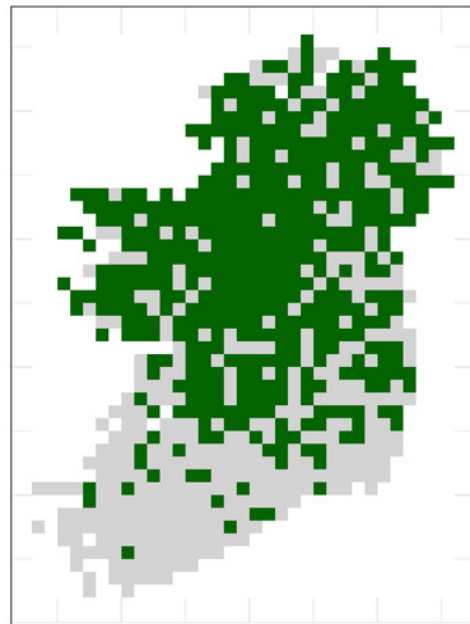



Figure 5

The number of virtual species successfully modeled.

The number of virtual species (out of 110 total species chosen for modelling from the large community simulation) for which species distribution models fitted within the computation time and memory constraints we imposed, according to the spatial sampling bias and sample size of training data and the species distribution modelling method. Species distribution modelling methods were (A) generalized linear models, (B) boosted regression trees, and (C) inverse distance-weighted interpolation. Spatial biases were no bias (Simpson evenness = 1), low (e.g. birds, Simpson evenness = 0.76), median (e.g. butterflies, Simpson evenness = 0.13), and severe (e.g. moths, Simpson evenness = 0.02).

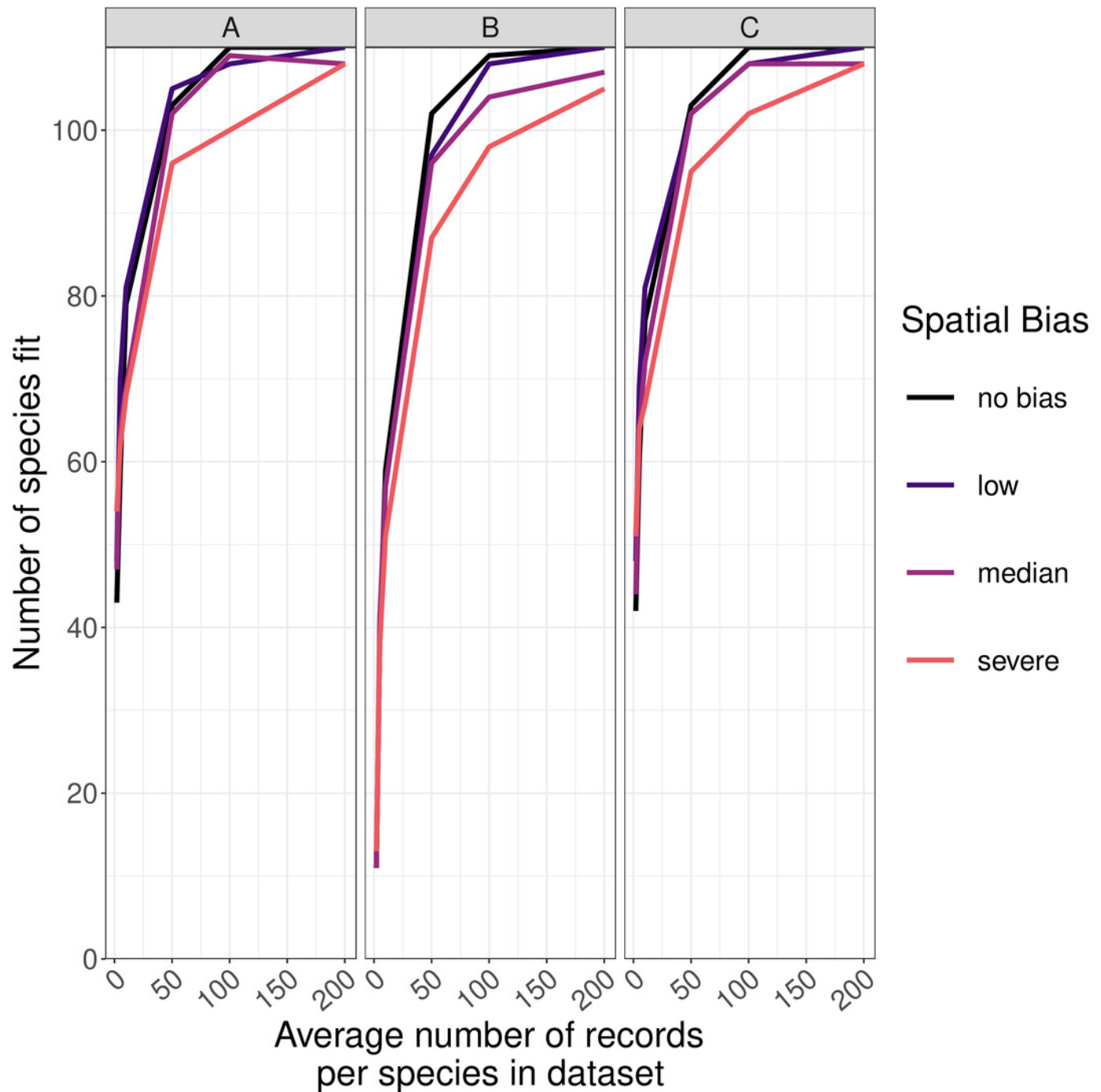



Figure 6

Expected prediction performance of species distribution models for 110 simulated species under a range of sample size and spatial sampling bias scenarios.

Panels show the expected prediction performance of species distribution models constructed using (A) generalize linear models, (B) boosted regression trees, and (C) inverse distance-weighted interpolation. Lines show expected area under the receiver operating characteristic curve (AUC) given the sample size and spatial sampling bias of training data, and the species distribution modelling method. Rug plots indicate sample sizes (mean number of records per species) of the virtual biological records datasets used to train species distribution models.

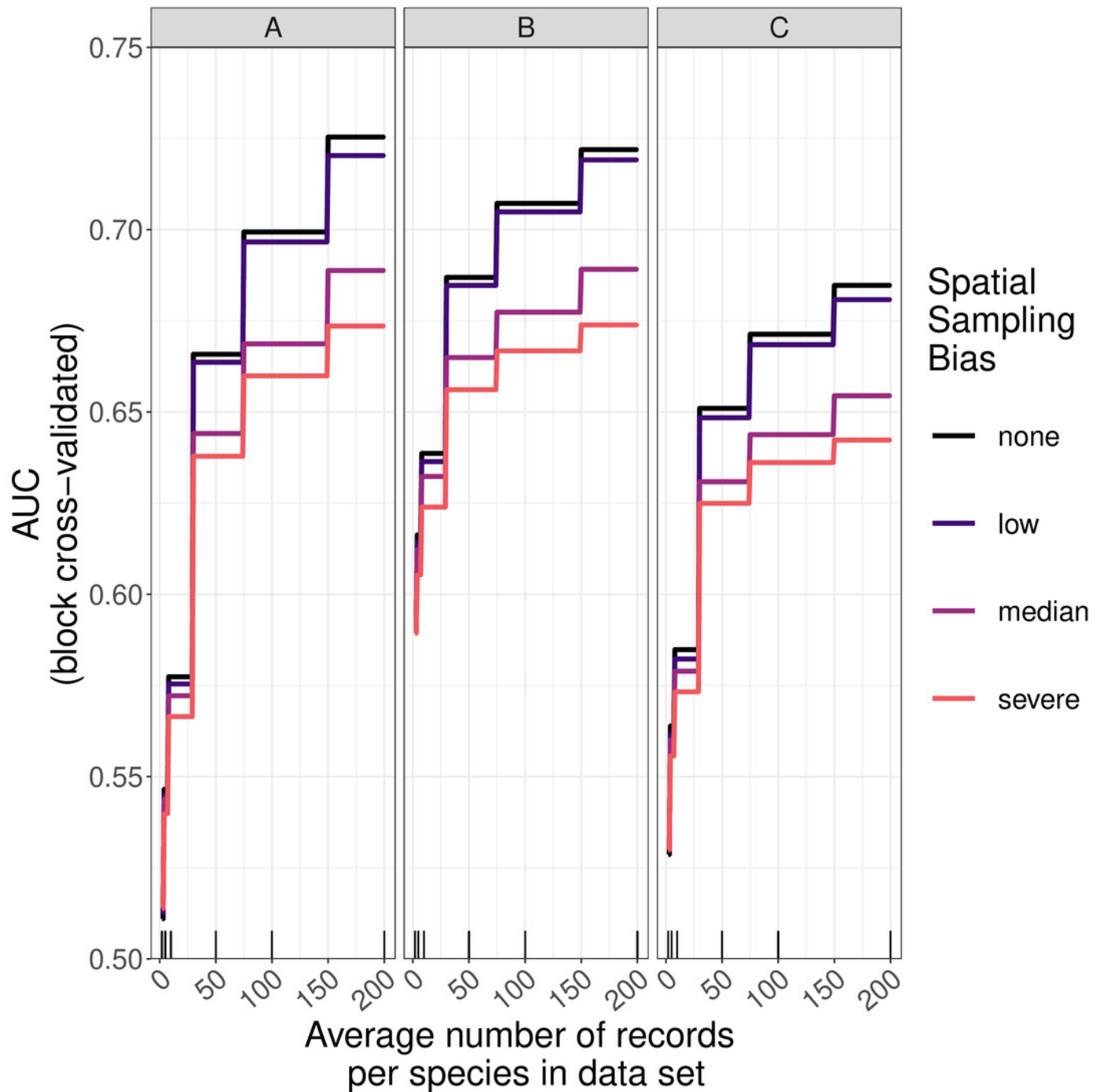



Figure 7

Observed prediction performance (AUC) of species distribution models for 110 virtual species under a range of sample size and spatial sampling bias scenarios.

Panels show the observed area under the receiver operating characteristic curve (AUC) of species distribution models constructed using (A) generalized linear models, (B) boosted regression trees, and (C) inverse distance-weighted interpolation. Boxes contain the middle 50% of the observed AUC values. The horizontal line within each box indicates the median AUC value. Each box plot (box, whiskers, and outlying points) represents 110 observations (one for each virtual species) unless models failed to fit for some species (see Fig. 4). The width of boxes is proportional to the square root of the number of observations in that group.

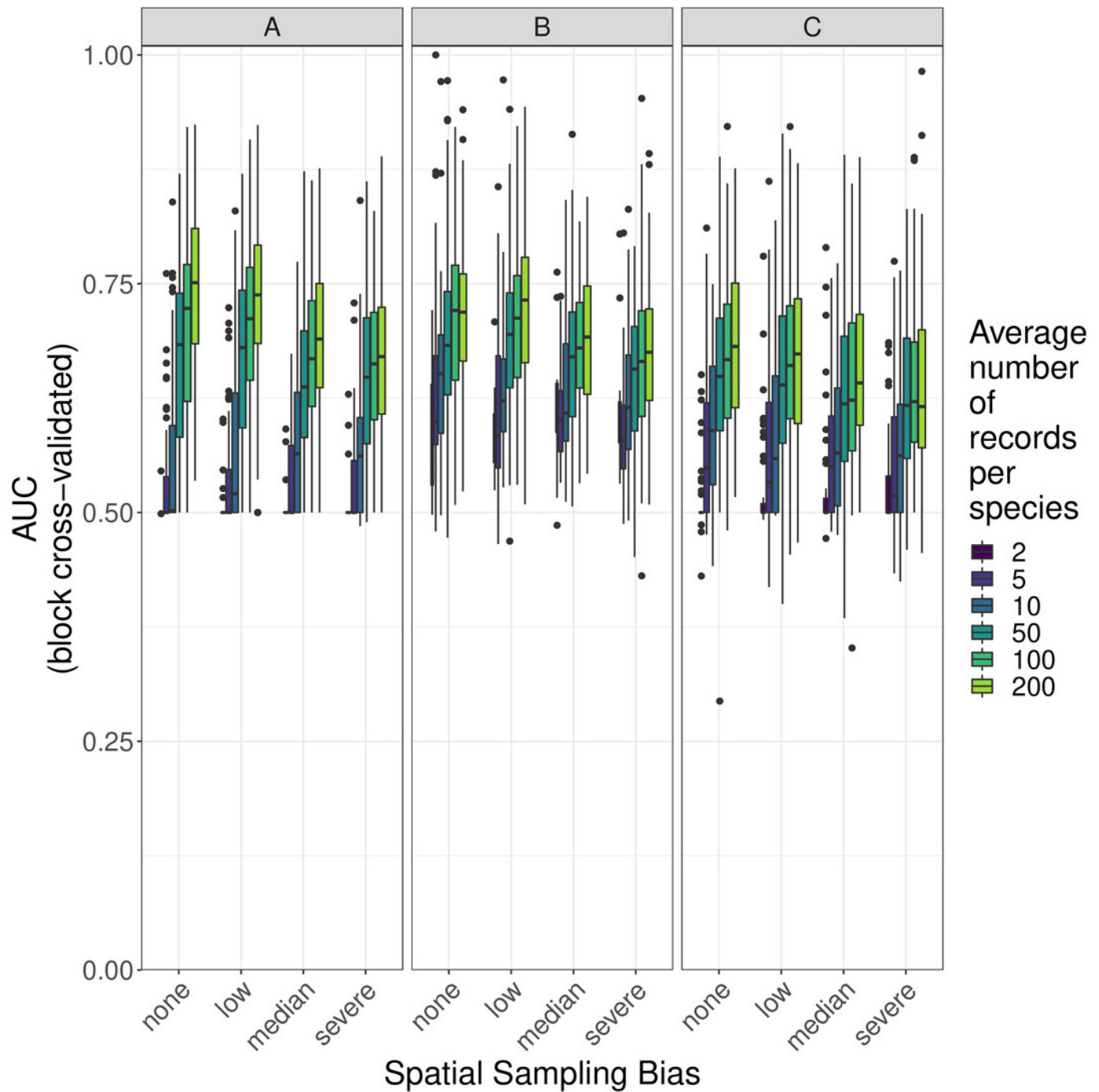



Figure 8

Observed prediction performance (RMSE) of species distribution models for 110 virtual species under a range of sample size and spatial sampling bias scenarios.

Panels show the observed root mean squared error (RMSE) of species distribution models constructed using (A) generalized linear models, (B) boosted regression trees, and (C) inverse distance-weighted interpolation. Boxes contain the middle 50% of the observed RMSE values. The horizontal line within each box indicates the median RMSE value. Each box plot (box, whiskers, and outlying points) represents 110 observations (one for each virtual species) unless models failed to fit for some species (see Fig. 4). The width of boxes is proportional to the square root of the number of observations in that group.

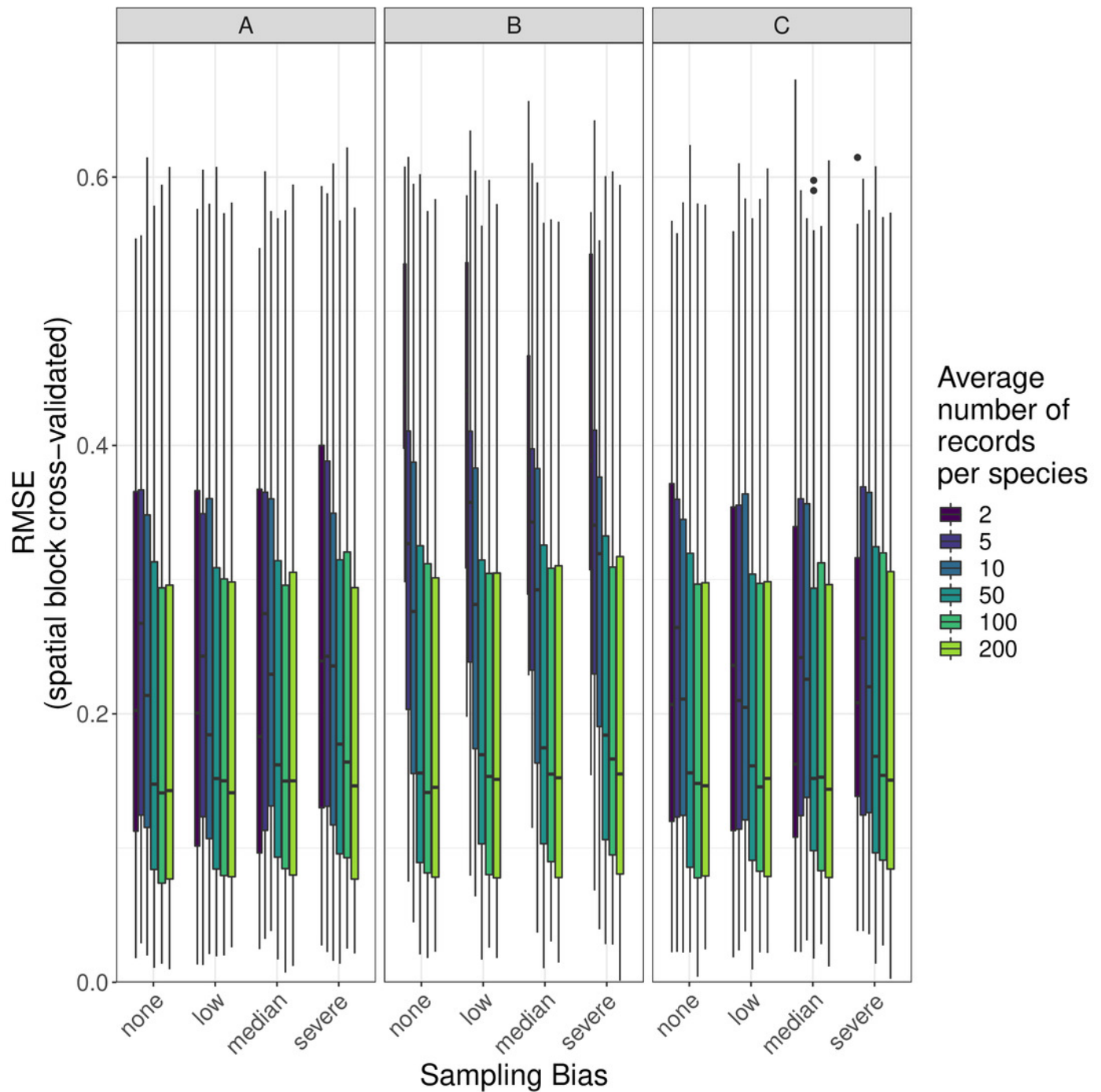
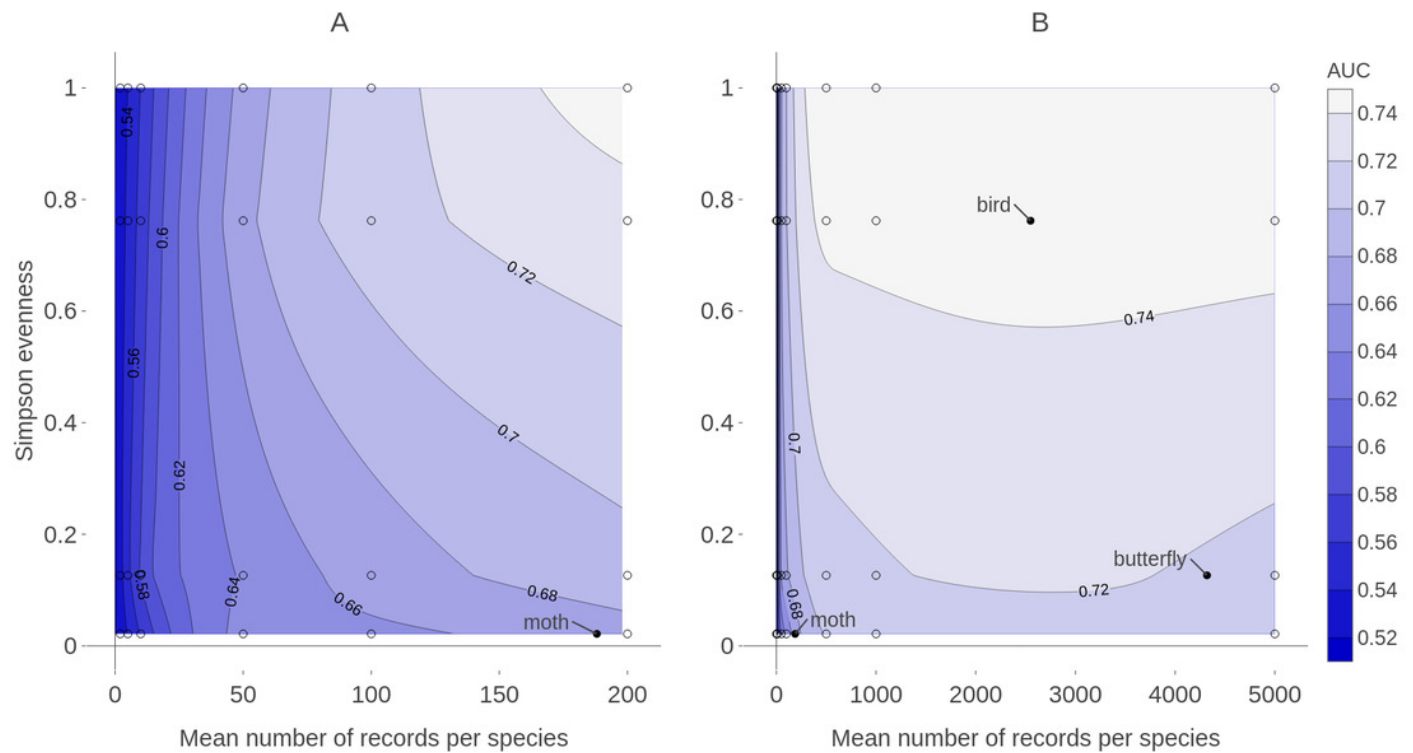



Figure 9

Contour plot of expected prediction performance of species distribution models as a function of the sample size and spatial sampling bias in virtual biological records datasets.

Expected prediction performance (AUC, contours and shading) of generalized linear model (GLM) species distribution models for virtual species, according to the spatial sampling evenness and sample size of training data. Note the different scales of the horizontal axes in A and B. (A) shows detail of prediction performance changes when sample sizes were small. (B) shows larger sample sizes. Spatial sampling evenness was quantified using Simpson evenness. High values of Simpson evenness indicate minimal spatial bias. Open circles show the values of sample size and spatial sampling evenness for virtual biological records datasets used to train species distribution models. Filled black circles show sample size and spatial sampling evenness of Irish biological records datasets used as spatial sampling templates.

Environmental predictor variables used to define and model the distribution of virtual species in Ireland.

Moran's I values indicate the spatial clustering of values for each variable, where a value of one indicates strong spatial clustering of variable values, zero indicates random spatial arrangement of values, and negative one indicates strongly dispersed spatial arrangement of values. Details of data sources are in Section 2.1.

Variable	Description	Data Source	Moran's I
annual minimum temperature (degrees C)	2% quantile of annual temperatures in each grid cell averaged over the years 1995-2016	E-OBS	0.84
annual maximum temperature (degrees C)	98% quantile of annual temperatures in each grid cell averaged over the years 1995-2016	E-OBS	0.83
annual precipitation (mm)	Average total annual precipitation in each grid cell over the years 1995-2016 (excluding 2010-2012)	E-OBS	0.82
average daily sea level atmospheric pressure (hecto Pascals)	Average daily sea level atmospheric pressure over the years 1995-2016	E-OBS	0.86
agricultural areas	Proportion of each grid cell classified as agricultural areas	CORINE Land Cover Database	0.53
artificial surfaces	Proportion of each grid cell classified as artificial surfaces	CORINE Land Cover Database	0.44
forest and semi-natural areas	Proportion of each grid cell classified as forest and semi-natural areas	CORINE Land Cover Database	0.41
water bodies	Proportion of each grid cell classified as water bodies	CORINE Land Cover Database	0.35
wetlands	Proportion of each grid cell classified as wetlands	CORINE Land Cover Database	0.55
elevation	Average elevation in each grid cell	ETOPO1	0.29

Spatial sampling evenness of the spatial sampling template datasets measured using Simpson evenness.

A value of one indicates perfectly even sampling (all grid squares containing the same number of records). Lower Simpson evenness values indicate more spatially uneven sampling.

Spatial sampling template	Simpson evenness value
no bias	1
low bias (birds)	0.762
median bias (butterflies)	0.126
severe bias (moths)	0.021

Importance of sample size, spatial bias, modelling method, and species prevalence for determining predictive performance of species distribution models.

Variable importance measures from a boosted regression tree show the relative influence of sample size (average number of records per species), species distribution modeling method, and spatial bias in training data on three different measures of prediction performance of species distribution models. The relative influence for each variable is the reduction in squared error attributed to that variable in a boosted regression tree model. Prediction performance metrics were area under the receiver operating characteristic curve (AUC), root mean squared error (RMSE) and Cohen's Kappa calculated using the threshold that maximized Kappa (Kappa). The effect of species prevalence was not included for AUC and for Kappa, because exploratory plots showed no indication of an effect.

Prediction performance metric	Variable	Relative importance (reduction in squared error)
AUC	Average number of records per species	78.5
	Species distribution modelling method	14.8
	Spatial bias	6.7
RMSE	Species prevalence	99.9
	Average number of records per species	0.2
	Species distribution modelling method	0.003
	Spatial bias	0.0001
Kappa	Species distribution modelling method	74.2
	Average number of records per species	22.4
	Spatial bias	3.4