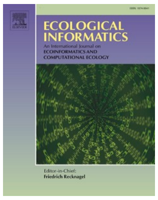
Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



A data-integration approach to correct sampling bias in species distribution models using multiple datasets of breeding birds in the Swiss Alps

Nasrin Amini Tehrani[a,*], Babak Naimi[b], Michel Jaboyedoff[a]

[a] Faculty of Geosciences and Environment, University of Lausanne, CH-1015 Lausanne, Switzerland
[b] Department of Geosciences and Geography, University of Helsinki, PO Box 64, 00014 Helsinki, Finland

ARTICLE INFO

Keywords:
Bird species
Data pooling
Model-based data integration
Presence-only data
Random forest
Species distribution models
Swiss Alps

ABSTRACT

It is essential to accurately model species distributions and biodiversity in response to many ecological and conservation challenges. The primary means of reliable decision-making on conservation priority are the data on the distributions and abundance of species. However, finding data that is accurate and reliable for predicting species distribution could be challenging. Data could come from different sources, with different designs, coverage, and potential sampling biases. In this study, we examined the emerging methods of modelling species distribution that integrate data from multiple sources such as systematic or standardized and casual or occasional surveys. We applied two modelling approaches, “data-pooling” and “ model-based data integration” that each involves combining various datasets to measure environmental interactions and clarify the distribution of species. Our paper demonstrates a reliable data integration workflow that includes gathering information on model-based data integration, creating a sub-model of each dataset independently, and finally, combining it into a single final model. We have shown that this is a more reliable way of developing a model than a data pooling strategy that combines multiple data sources to fit a single model. Moreover, data integration approaches could improve the poor predictive performance of systematic small datasets, through model-based data integration techniques that enhance the predictive accuracy of Species Distribution Models. We also identified, consistent with previous research, that machine learning algorithms are the most accurate techniques to predict bird species distribution in our heterogeneous study area in the western Swiss Alps. In particular, tree-dependent ensembles of Random Forest (RF) contribute to a better understanding of the interactions between species and the environment.

1. Introduction

An effective conservation planning relies on reliable information on biodiversity distribution that has been widely available in the forms of species occurrences (Brooks et al., 2006; Carvalho et al., 2010; Liu et al., 2013). Changes in species distributions and biodiversity have profound consequences for wildlife and species conservation planning (Brotons et al., 2007; Carvalho et al., 2011; Wilson et al., 2004). Confidence in conservation decisions depend on the type of data used in Species Distribution Models (SDMs) and most practical decisions mainly address the issue of how much data on species distribution is available (Carvalho et al., 2010; Hortal et al., 2007; Tulloch et al., 2016). Nonetheless, most species data documentation is still incomplete, if not unavailable (Araujo and Guisan, 2006; Braunisch and Suchant, 2010; Pressey, 2004). In particular, systematic surveys on species data are often not feasible in large geographic areas, restricting conservationists and environmentalists to effectively use such incomplete and imprecise species data (Braunisch and Suchant, 2010; Fajardo et al., 2014; Niel and Lebreton, 2005).

The occurrence (presence/absence) of species at various sites is normally associated with specific environmental variables, necessary to understand habitat preferences of species, predict their distributions, and inform conservation decisions (Ferrier et al., 2002; Guillera-Arroita et al., 2015; Rodríguez et al., 2007; Smeraldo et al., 2020; Wagner et al., 2020). While researchers continue to use SDMs with imperfect presence-only data, mostly derived from opportunistic sampling based on museum collections, biological inventories, or citizens’ science, using such data may have some consequences in the models (Carvalho et al., 2010; Dickinson et al., 2010; Fithian et al., 2015; Fletcher Jr et al., 2019). Models with inappropriate data could waste valuable resources

* Corresponding author.
E-mail address: Nasrin.AminiTehrani@unil.ch (N.A. Tehrani).

https://doi.org/10.1016/j.ecoinf.2021.101501
Received 3 May 2021; Received in revised form 23 November 2021; Accepted 24 November 2021
Available online 29 November 2021

and produce results that are unlikely to address the problem at hand (Guillera-Arroita et al., 2015; Hernandez et al., 2006; Wilson et al., 2005). In the case of SDMs, they may perform differently depending on the source of data used to fit the models that may be systematic, standardized data sets within a restricted area, or "volunteer-based monitoring schemes" (VMS) collected imprecisely at a wide regional level with varying spatial resolution (Braunisch and Suchant, 2010; Dickinson et al., 2010; Isaac et al., 2020; Ratnieks et al., 2016; Steen et al., 2019).

Although data integration from various sources is becoming increasingly popular as a potential solution, it can also be challenging (Carvalho et al., 2010; Fletcher Jr et al., 2019; Isaac et al., 2020) since each data source has major variations in assumptions, design, environmental coverage, and potential sampling biases (Bird et al., 2014; Fletcher Jr et al., 2019; Pacifici et al., 2017). In SDMs, it is therefore important and challenging to properly integrate the various data sources so that predictions and statistical deductions are more precise (Dorazio, 2014; Fithian et al., 2015; Pacifici et al., 2017; Talluto et al., 2016). Where there is a lack of data available for a given species, the data integration approach would be able to account for various sampling biases affecting the data and improve the models' predictive performance by increasing the quantity of available data and optimizing the useful information for predicting species distributions (Fithian et al., 2015; Fletcher Jr et al., 2019; Isaac et al., 2020).

In addition to the data, there can also be uncertainties in the predictive functions of the algorithms (modelling techniques) that apply to connect the occurrence data (dependent data) to the independent variable (environmental variables) (Moudrý and Šímová, 2012; Pearson et al., 2006; Watling et al., 2015). The most significant source of uncertainty in SDMs' performance and spatial predictions is the choice of modelling algorithm as each algorithm approaches the interaction between species occurrence and environment in different ways (Watling et al., 2015). As a result, deciding on the most effective SDM algorithm is challenging, as they are numerous and useful in many ways (Aguirre-Gutiérrez et al., 2013; Dormann et al., 2008; Elith et al., 2006; Watling et al., 2015). It is, therefore, necessary to find out which modelling approach performs better than others as applying inappropriate modelling technique could output an overestimation of species distributions, affect environmental planning decisions negatively, and waste resources (Carvalho et al., 2010; Elith et al., 2006; Mendes et al., 2020; Segurado and Araujo, 2004; Watling et al., 2015).

To achieve complementary advantages from different sources of data, some studies suggested a probabilistic model framework that integrate the presence-only and survey data from multiple sources. They pooled presence-only and presence-absence data and maximized joint likelihoods, calculating and correcting concurrently sampling bias impacting the presence-only data. They discovered that the data-pooling methodology significantly improves the model's out-of-sample predictive efficiency where there is a lack of available presence-absence data for a given species (see Fithian et al., 2015). Some developed a method of three modelling strategies of 'shared,' 'correlation,' and 'covariates' for the jointly modelling of two data sources (one of high quality and one of lesser quality) (Pacifici et al., 2017). The findings showed that all three of the methods which used the secondary data source could optimize out-of-sample estimates compared to a single data source. They proposed that these approaches are robust alternatives when nothing is known about secondary data obtained opportunistically or by citizen scientists (see Pacifici et al., 2017). Other research involved integrating a variety of data sets with different sampling methodologies and amounts of data for species distribution. They highlighted the effects of data integration and how it can improve the accuracy in environmental relations, predictive performance, and address sample biases (see Fletcher Jr et al., 2019).

In this paper, we compared the model performance using eight different bird datasets which were collected in systematic or casual ways. Additionally, we examined model performance in which different types of data were integrated in either a data-pooling process or model-based integration. Our objectives then would be, (a) is a data pooling technique better at modelling species distributions than just using a systematic or casual data set? And (b) which modelling algorithm performs better with each dataset? Most studies compare findings from different techniques to similar data sets (Marmion et al., 2009; Parviainen et al., 2009; Pearson et al., 2006; Thuiller, 2003). However, it is essential to study the relative output of various modelling techniques through various data sources since it is an ongoing problem in ecology and conservation biology that demands more research. We propose the integration of different bird data sources with different modelling techniques to make wider and more rigorous use of species distribution information in wildlife conservation, both by raising the overall quality of the data and by expanding the number of species with sufficient information to be included in spatial analyses (Fithian et al., 2015; Merow et al., 2017; Zhang and Vincent, 2017). The interest, therefore, has emerged in developing different techniques for integrating different types of datasets to improve the estimation and comprehensive descriptions of potential and realized species distributions in space and time (Fithian et al., 2015; Fletcher Jr et al., 2019; Isaac et al., 2020; Miller et al., 2019; Pacifici et al., 2017).

2. Material and methods

2.1. *Study area and species data*

The study area is in the western Swiss Alps in Vaud (46°10′ to 46°30′N; 6°50′ to 7°10′E; Fig. 1). Since 2013, it has been an interdisciplinary and transdisciplinary research site for the University of Lausanne, and it currently belongs to the Interdisciplinary Centre for Mountain Studies (CIRM) (http://rechalp.unil.ch). It covers an area of approximately 700 km^2 and an elevational gradient extending from Lake Geneva at 372 m a.s.l. to Pointe des Diablerets at 3210 m a.s.l. (Amini Tehrani et al., 2020, 2021; Descombes et al., 2017; Scherrer et al., 2019). Due to anthropogenic activities such as the dense population and intensive farming in the Rhône Valley, tourism and outdoor activities, and more extensive farming in the subalpine regions, this area has been dominated by a mosaic of meadows, pastures, and forest and woodland patches (see http://rechalp.unil.ch; Randin et al., 2009; Scherrer et al., 2019; Amini Tehrani et al., 2020, 2021). The Swiss Ornithological Institute (Monitoring Häufige Brutvögel [MHB]; Schmid et al., 2004) provided us with the bird data (presence-only), which has been recorded annually since 1999 (for more information on the survey, see https://www.vogelwarte.ch/de/projekte/monitoring/monitoring-haeufige-brutvoegel).

Based on the sampling strategy, we classified the data into two categories: systematic and standardized surveys, and casual and occasional ones (Fig. 2, Table 1) (more information in supplementary).

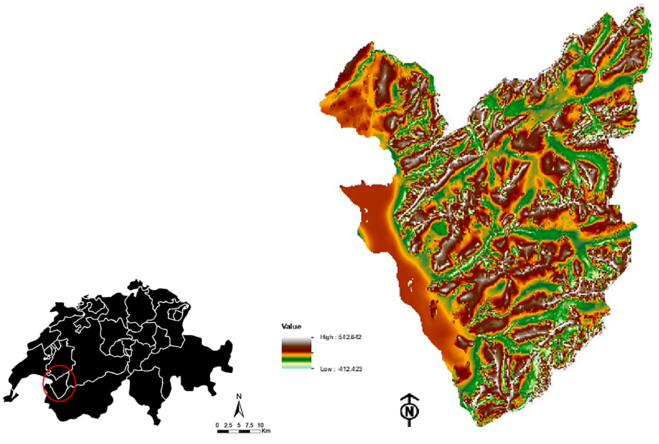


Fig. 1. Study area in the western Swiss Alps.

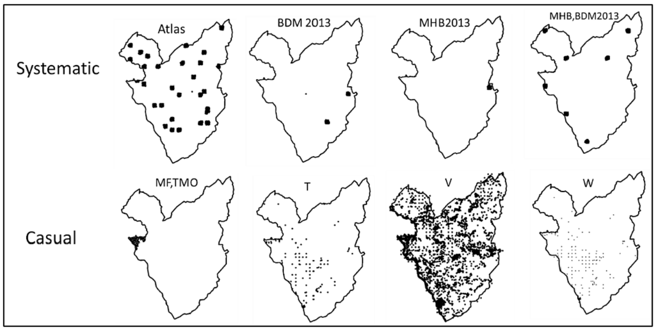


Fig. 2. Sampling points (presence points) for each dataset across the study area. The number of sites for each dataset is: Atlas data = 12,580, BDM 2013 = 712, MHB 2013 = 658, MHBBDM 2013 = 13,661, MF, TMO = 1204, T = 371, V = 14,290, W = 2272.

2.2. Mixed data sets

We created two “mixed data sets” (Ghysels et al., 2004) because statistical models developed upon the bird datasets with small sample size and potential biases may become unreliable (if data are not large enough and are not well-collected to cover all environmental gradients in the study area) (Abadi et al., 2010; Moudrý and Šímová, 2012).

2.2.1. Mixed data set 1 (107 species)

Mixed data set 1 combines all systematically recorded data including “BDM ≥ 2013”, “MHB ≥ 2013”,” MHB + BDM ≥ 2013″, and “MF, TMO” (Fig. 3). We mixed these bird data sets since they were recorded systematically in squares of 1*1 km and were supported by the online tool «Terri map online» which applies the standardization process and help avoid mistakes (Birrer, 2019) (see Table 1 for more information on the survey https://www.vogelwarte.ch/assets/files/projekte/ueberwachung/datenabgabe/Directives%20fz%202019.pdf).

2.2.2. Mixed data set 2 (177 species)

We combined three casual bird datasets, “T” (complete observation list), “V” (single observation), and “W” (complete observation list) which were recorded using non-systematic methods (Table 1). As step 1

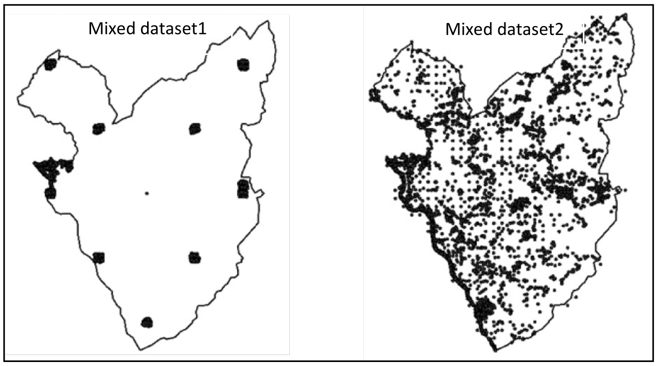


Fig. 3. Combination of all systematic data (Mixed dataset1) and casual data (Mixed dataset2).

Table 1
Bird data collected using different strategies (systematic and casual).

Category	Name	Description	Remarks	Number of bird species in each group of data	Geography precision ID	Description
Systematic	Atlas 2013-2016	Swiss Breeding Bird Atlas 2013–2016		79	1–3	1 precise per meter (precise per hectare in case of aggregation) 3 flying bird
	BDM ≥ 2013	Swiss Biodiversity Monitoring	Since 2013 surveys are obtained through the online tool “Terrimap Online” for 1x1km2 surveys	23	1–5	1 precise per meter 5 inside this 1 × 1 km square
	MHB ≥ 2013	Common Breeding Bird Survey	Since 2013 surveys are obtained through the Online tool “Terrimap Online”	9	1–3	1 precise per meter (precise per hectare in case of aggregation) 3 flying bird
	MHB + BDM ≥ 2013	combination of BDM and MHB data, Monitoring of widespread breeding birds + Swiss biodiversity monitoring	Since 2013 surveys are obtained through the Online tool “Terrimap Online”	60	1–3	1 precise per meter (precise per hectare in case of aggregation) 3 flying bird
	MF, TMO	Wetland monitoring	Terrimap Online (online tool for 1x1km2 surveys)	15	1	1 precise per meter (precise per hectare in case of aggregation)
Casual	T	Complete observation list		14	1-4-5	1 precise per meter 4 data attributed to the nearest locality (precision unknown) 5 inside this 1 × 1 km square
	V	Single observation		108	1-4-5	1 precise per meter 4 data attributed to the nearest locality (precision unknown) 5 inside this 1 × 1 km square
	W	Complete observation list		55	1-4-5	1 precise per meter 4 data attributed to the nearest locality (precision unknown) 5 inside this 1 × 1 km square

of the analysis, we integrated all the eight bird data sets separately including both systematic ("Atlas 2013–2016", "BDM $\geq$ 2013", "MHB $\geq$ 2013"," MHB + BDM $\geq$ 2013″ and casual ones ("MFTMO", "T", "V", and "W"). These datasets are all presence-only data, meaning they only contain presence records (Fig. 3).

2.3. Species distribution modelling

2.3.1. Step 1: initial evaluation of predictive accuracy of eight bird datasets and improvement of the datasets with unreliable model evaluations values

We used species distribution models across eight datasets (Atlas 2013–2016, BDM 2013, MHB 2013, MHBBDM 2013, MFTMO, T, V, and W) to better understand the interactions between sampling methods, the various types of bird species data, modelling techniques, and modelling accuracy (Edwards Jr et al., 2006; Guisan and Thuiller, 2005; Jiménez-Valverde et al., 2009; Parviainen et al., 2009). For each dataset, we fitted SDMs using different environmental variables at 100 m spatial resolution (more information in supplementary) and manipulated them in ARCGIS 10.2 (Environmental System Research Institute, Inc.) or in *R*3.3 (R Core Team, 2016) because each dataset has a different geographic distribution area (Fig. 2) and distribution of sample location (distribution of species) has an important effect on model accuracy (Brotons et al., 2007; Hirzel and Guisan, 2002). Therefore, for each dataset, we selected different sets of environmental variables from various sources (Gama et al., 2016; Parra et al., 2004), assumed to be ecologically meaningful and influential on bird species (see more details about environmental variables in supplementary information). We tested the pairwise correlations for all predictors (Table S1) to derive the simplest probable distributions for each dataset based on uncorrelated variables (Spearman correlation >0.7; Dormann et al., 2013) and reduce collinearity.

We applied eight modelling techniques to each of the eight data sets to identify the most important modelling techniques with the highest predictive accuracy (Barbet-Massin et al., 2012; Grenouillet et al., 2011; Marmion et al., 2009; Parviainen et al., 2009). Modelling techniques include Flexible Discriminant Analysis (FDA; Grenouillet et al., 2011; Reiss et al., 2011), Classification Tree Analysis (CTA; Marini et al., 2010), Multivariate Adaptive Regression Splines (MARS) (Marini et al., 2010), Generalized Linear Models (GLM) (Bikkina, 2014), Random Forest (RF) (Tonini et al., 2020), Artificial Neural Networks (ANN) (Manel et al., 1999), Generalized Boosting Model (GBM) also known as Boosted Regression Trees/BRT (Heikkinen et al., 2012), and Surface Range Envelop (SRE) equivalent to BIOCLIM (Barbet-Massin et al., 2012; Booth et al., 2014; Thuiller et al., 2016). We compared all these SDM techniques using different environmental variables for each group of the bird dataset (More information in Supplementary). We employed the biomod2 package (Thuiller et al., 2016) in R v3.3 (R Core Team, 2016) to fit the species distribution models across the eight bird datasets.

For each dataset, we ran SDMs 10 times with 5 replications of pseudo-absence records with a size of 10,000 (Barbet-Massin et al., 2012), for a total of 50 runs. At each run, we selected 70% of the records at random to train the model. For the remaining 30%, we used an evaluation dataset (a combination of the systematic datasets MHB, BDM, and MHBBDM) with high precision geography identification (Table 1) from the similar study area to evaluate the models (Edwards Jr et al., 2006; Graham et al., 2008; Hallman and Robinson, 2020). For model evaluations, we used several indices such as the area under the receiver operating characteristic curve (AUC; Jiménez-Valverde, 2012; Fernandes et al., 2019), the true skills statistic (TSS; Allouche et al., 2006; Fernandes et al., 2019), and Cohen's Kappa Statistic (KAPPA; Cohen, 1960; Fernandes et al., 2019) (Li et al., 2020; Smeraldo et al., 2021).

We applied equal weighting to pseudo-absence and presence records to arrive at an overall prevalence of 0.5 (Amini Tehrani et al., 2020; Ferrier et al., 2002; Scherrer et al., 2019; Thuiller et al., 2016) because the accuracy of the models is reduced by unbalanced prevalence (Guisan et al., 2017). We selected the best values of model evaluation indices for each dataset based on the highest model evaluation scores. Then, we identified the datasets with the reliable model evaluation values to be used later in step 2 (Data-pooling technique) and step 3 (Model-based data integration).

2.3.2. Data combining in species distribution

Integrated SDMs, models that simultaneously combine different data sources on species locations to quantify environmental relationships that contribute to the understanding of species distribution, tackled many of the data integration challenges (Carvalho et al., 2010;Fithian et al., 2015; Fletcher Jr et al., 2019; Isaac et al., 2020). Data integration could increase the quantity of available data and contribute to enhanced predictions of species distributions (Miller et al., 2019; Pacifici et al., 2017). We identify two ways in which multiple sources of data are typically combined for modelling species distributions.

2.3.3. Step 2: data-pooling technique

Data pooling refers to a common method that incorporates observations (presence-only data) from different sources, regardless of data source and/or sampling issues (Barbet-Massin et al., 2012; Fithian et al., 2014, 2015; Tsoar et al., 2007) that could be based on a single observation model that may fail to identify the specification of data sources, ignore their discrepancies, or reduce data to a standard in the model (Fithian et al., 2015; Fletcher Jr et al., 2019; Isaac et al., 2020). It implies that all variations between datasets are minimal enough which could be discarded or lowered to a lowest common denominator (Isaac et al., 2020). This method could boost the out-of-sample predictive performance of the model if the data available for a species are insufficient (Fithian et al., 2014, 2015; Fletcher Jr et al., 2019; Saracco et al., 2008).

We applied the data pooling method to the bird datasets that had small distribution across the study area and were not technically suitable for making reliable SDM (Fithian et al., 2015; Fletcher Jr et al., 2019; Picard et al., 2009). They were obtained from the step 1 of the analysis (initial evaluating predictive accuracy). These datasets included systematic datasets of "BDM $\geq$ 2013", "MHB $\geq$ 2013"," MHB + BDM $\geq$ 2013″, and casual ones "MF, TMO". All these datasets are pooled and combined to make 'mixed data set 1' (107 species) (Fig. 2). We included "MF, TMO" data in "mixed data set 1" because these data are supported by the online tool «Terrimap online» that applies the standardization process and helps avoid mistakes (Birrer, 2019). Therefore, it could be a reliable dataset like systematic data. To make "mixed data set 2" (177 species), we used a combination of three casual bird data "T" (complete observation list), "V" (single observation), and "W" (complete observation list) that were collected in a non-systematic strategy (Fig. 3). We built SDMs separately for the final datasets 'mixed data set 1′ (107 species) and 'mixed data set 2' (177 species) in the "biomod2" (Thuiller et al., 2016) in *R*3.3 software (R Core Team, 2016) with uncorrelated variables as predictors (see supplementary). Much like step 1 of the analysis, we applied eight modelling techniques (FDA, CTA, MARS, GLM, RF, ANN, GBM, SRE) separately to each of the two mixed data sets. We calibrated and evaluated all models by techniques identical to those in the previous step (Initial evaluation of predictive accuracy of eight bird datasets and improvement of the datasets with unreliable model evaluations values).

2.3.4. Model-based data integration

Model-based data integration (Isaac et al., 2020; Merow et al., 2017) or integrated SDMs (ISDMs) (Fletcher Jr et al., 2019; Schank et al., 2019; Simmonds et al., 2020) has been developed by integrating data sets in a way that the strengths of each dataset are retained (Isaac et al., 2020). Model-based data integration has its strength in spreading parameters across sub-models, which offers more reliable estimation of demographic parameters in comparison to the independent models (Fletcher Jr et al., 2019; Isaac et al., 2020; Miller et al., 2019). We applied a model-based data integration method to each systematic bird dataset "BDM $\geq$ 2013", "MHB $\geq$ 2013", "MHB + BDM $\geq$ 2013", and casual “MF, TMO”. We made a single SDM model (small model) for each data set and then ensembled all predictions of the small models across all bird datasets based on the average AUC. We repeated this process (making a small model for each dataset and then ensembling predictions across all bird data) for each casual bird dataset “T” (Complete observation list), “V” (single observation) and “W” (complete observation list). We applied eight modelling techniques (FDA, CTA, MARS, GLM, RF, ANN, GBM, SRE) to each of the data set separately to find the most important modelling techniques (data-pooling technique). We calibrated and evaluated all models using techniques identical to those in the previous step (data-pooling technique).

3. Results

3.1. Step 1: initial evaluation of predictive accuracy of eight bird datasets and improvement of the datasets with unreliable model evaluations values

Model accuracies were estimated for the three evaluation techniques (ROC, KAPPA and TSS) by taking average across the eight technique models and showed considerably different across the eight groups bird datasets. Data sets “MHB2013” ROC (0.98), kappa (0.94), TSS (0.95), “MF-TMO” ROC (0.98), KAPPA (0.93), TSS (0.95), “MHB, BDM2013” ROC (0.88), KAPPA (0.71), TSS (0.70), “BDM2013” ROC (0.88), KAPPA (0.47), TSS (0.71), “W” ROC (0.81), KAPPA (0.45), TSS (0.54), “Atlas2013-16” ROC (0.75), KAPPA (0.41), TSS (0.42), “T” ROC (0.76), KAPPA (0.31), TSS (0.49), “V” ROC (0.74), KAPPA (0.40), TSS (0.42) showed higher prediction accuracy respectively across eight bird datasets (Fig. 4). There were also distinguishing differences in projections of the species distribution among the eight modelling techniques across eight bird datasets and the results showed that RF and GBM predict distribution of bird species in the study area better than other modelling techniques, MHB2013 (RF = 1, GBM = 0.99), MFTMO (RF = 0.98, GBM = 0.98), MHBBDM (RF = 0.96, GBM = 0.92), BDM2013 (RF = 0.80, GBM = 0.79), Atlas (RF = 0.76, GBM = 0.63), T (RF = 0.72, GBM = 0.66), V (RF = 0.61, GBM = 0.56), W (RF = 0.55, GBM = 0.63) (Fig. 5) except for W dataset that ANN (0.86) and CTA (0.74) had the highest AUC among other modelling techniques (Fig. 5). Overall, Random Forest algorithm was clearly the most effective modelling techniques for predicting bird distribution in the study area, according to our findings. The results obtained with the Random Forest are very encouraging for predicting habitat suitability, presenting the highest accuracy among the other algorithms tested in this study.

3.2. Step 2: data-pooling technique

Data-pooling approach, without regard for the data source, showed that “mixed data set 1” with ROC (0.81), KAPPA (0.55), TSS (0.54) (Figs. 6, 7) had a more dependable and unbiased model accuracy than that of the single systematic data (Figs. 4, 5). It could be the best bird dataset for predicting bird species distribution in the study area as compared to other datasets such ad “Atlas 2013–2016”, ROC (0.76), KAPPA (0.42), TSS (0.43) and “mixed data set 2” (177 species), ROC (0.74), KAPPA (0.40), TSS (0.40) (Fig. 6). The pooling approach, in contrast to the single dataset (W, T, V) (Fig. 4), could not improve model accuracy (ROC (0.74), KAPPA (0.40), TSS (0.40) (Fig. 6)) of the casual datasets like “mixed data set 2” (177 species) as the values of model evaluations decreased in combining and pooling data. The results also identified RF as the most important modelling technique across “mixed data set 1”, with ROC (0.99), KAPPA (0.91), TSS (0.91) and “mixed data set 2”, with ROC (0.81), KAPPA (0.53), TSS (0.52) respectively (Fig. 7).

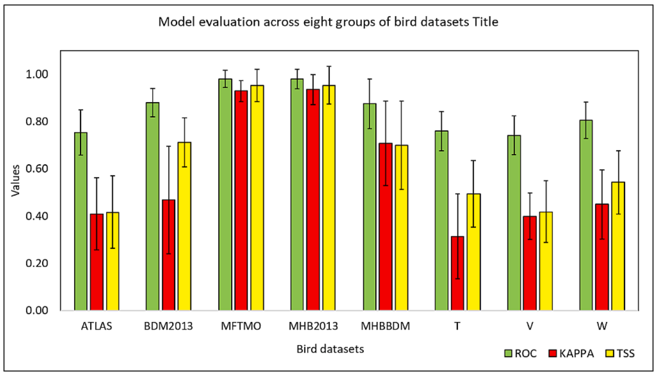


Fig. 4. Initial evaluation of predictive accuracy of the eight bird datasets based on receiver operating characteristic curve (ROC), KAPPA, True Skill Statistic (TSS), and improving dataset with low AUC (Area Under the Curve). See Table 1 for further information on each bird dataset.

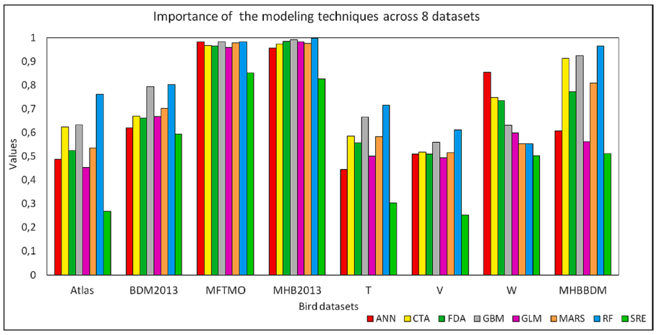


Fig. 5. Initial evaluation of predictive accuracy of the eight bird datasets across eight different algorithms: Artificial Neural Networks (ANN), Classification Tree Analysis (CTA), Flexible Discriminant Analysis (FDA), Generalized Boosting Model (GBM), Generalized Linear Models (GLM), Multivariate Adaptive Regression Splines (MARS), Random Forest (RF), Surface Range Envelop (SRE).

3.3. Step 3: model-based data integration

Model-based data integration of all systematic bird datasets (BDM ≥ 2013, MHB ≥ 2013) and casual datasets (MF, TMO), which involved constructing distinct data models for each source and then integrating the data, increased model accuracy with ROC (0.94), KAPPA (0.78), and TSS (0.91). (0.86) (Fig. 8). This is also true for casual dataset “T” (complete observation list), “V” (single observation), and “W” (complete observation list) with ROC (0.77), kappa (0.42), TSS (0.49) (Fig. 8). It outperformed the data-pooling strategy with ROC (0.74), KAPPA (0.40), TSS (0.40) (Fig. 6) in predicting species distribution. The findings of model-based data integration revealed that RF outperforms all other algorithms in terms of model evaluation: “mixed data set 1” with ROC (0.98), KAPPA (0.90), TSS (0.93), and “mixed data set 2” with ROC (0.88), KAPPA (0.64), TSS (0.66). Therefore, it could be considered as the most efficient algorithm (Fig. 9).

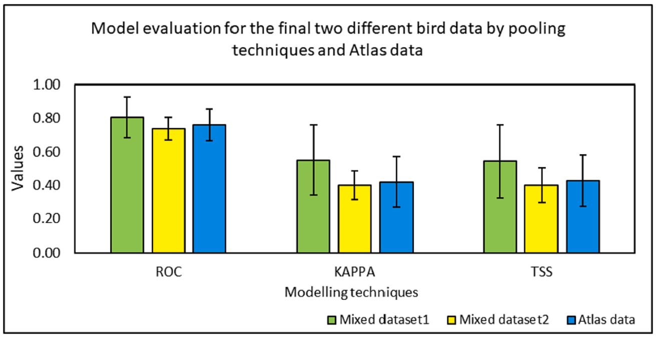


Fig. 6. Model evaluation for the final two different bird data (Mixed dataset 1, Mixed dataset 2) by pooling techniques and Atlas data.

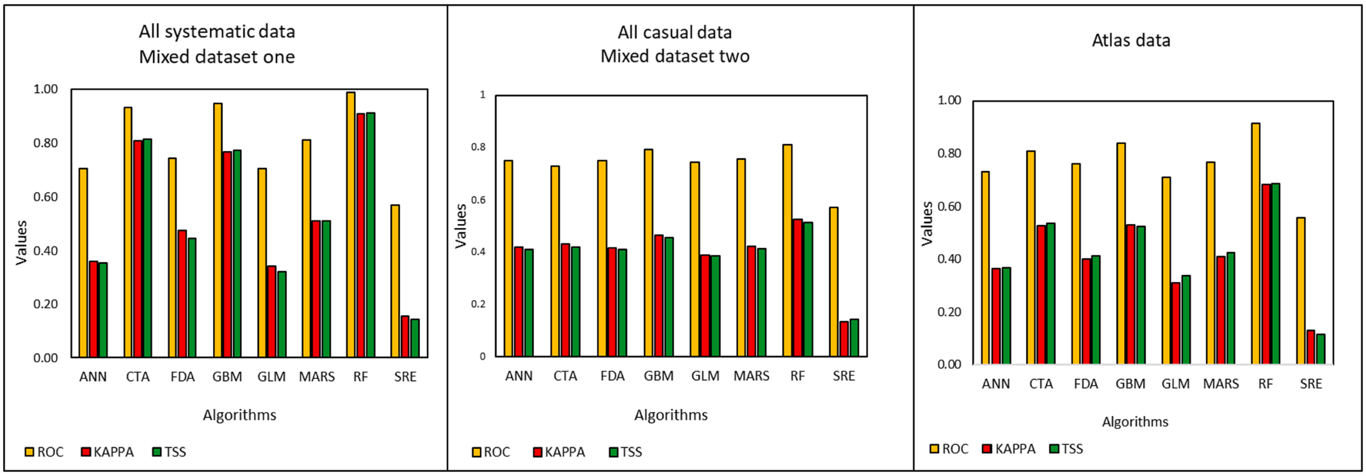


Fig. 7. Model evaluations by receiver operating characteristic curve (ROC), KAPPA, True Skill Statistic (TSS)) for the final two different bird data by pooling techniques and Atlas data across eight different algorithms (Artificial Neural Networks (ANN), Classification Tree Analysis (CTA), Flexible Discriminant Analysis (FDA), Generalized Boosting Model (GBM), Generalized Linear Models (GLM), Multivariate Adaptive Regression Splines (MARS), Random Forest (RF), Surface Range Envelop (SRE)).

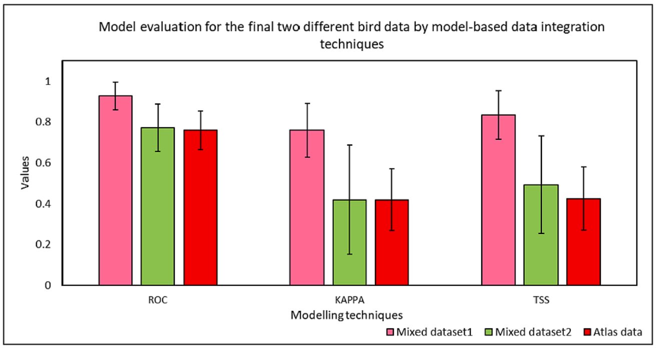


Fig. 8. Model evaluation for the final two different bird data using model-based data integration technique and Atlas data.

4. Discussions

In this study, we proposed two different methods that enabled the integration and combination of multiple sources of presence-only data to examine how they affect the performance of SDMs and then determine the most accurate bird dataset for predicting bird species distribution in the western Swiss Alps as the study area (Dorazio, 2014; Fithian et al., 2015; Miller et al., 2019). These methods illustrate how different types of data can be effectively combined in SDMs to produce accurate predictions of bird habitats suitability (Domisch et al., 2016; Koshkina et al., 2017). The combination of different sources of datasets showed that it can be appropriate especially if the quantity of data on species sites is limited in the planned surveys. Combining data could increase the sample size and consequently improve model accuracy (Fletcher et al., 2016; Fletcher Jr et al., 2019). We observed that SDMs can predict with greater accuracy and reliability when model-based data integration is applyed since this method takes into account the data obtained from different sampling designs (Fletcher Jr et al., 2019; Isaac et al., 2020; Merow et al., 2017).

We observed that models developed on a systematic collection of data with a small sample size performed poorly (with unreliable values of model evaluations). However, they can be enhanced by using data-integration approaches such as model-based data integration (Fithian et al., 2015; Isaac et al., 2020; Koshkina et al., 2017; Tenan et al., 2017). Model-based data integration approach is flexible because a broader variety of data types can be accommodated (Fletcher Jr et al., 2019; Isaac et al., 2020; Merow et al., 2017). It could clarify the differences in the integration of datasets hence retain their strengths and correct their weaknesses (Fletcher Jr et al., 2019; Isaac et al., 2020; Miller et al., 2019).

Our research methods provide an efficient solution for systematically collected data with a small sample size, allowing us to predict the distributions of birds more precisely and accurately in the mountainous area of the study (Fithian et al., 2015; Fletcher Jr et al., 2019). Here, a combined systematic bird dataset (mixed data set 1) could serve as a

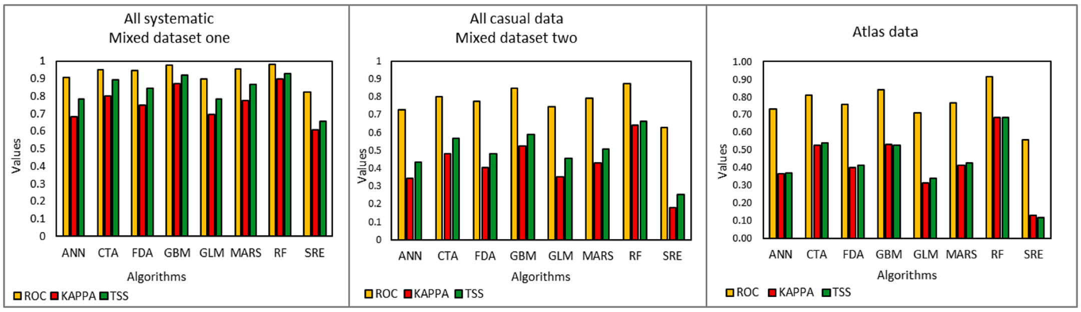


Fig. 9. Model evaluation by Receiver Operating Characteristic curve (ROC), KAPPA, True Skill Statistic (TSS) for the final two different bird data using model-based data integration technique and Atlas data across eight different algorithms (Artificial Neural Networks (ANN), Classification Tree Analysis (CTA), Flexible Discriminant Analysis (FDA), Generalized Boosting Model (GBM), Generalized Linear Models (GLM), Multivariate Adaptive Regression Splines (MARS), Random Forest (RF), and Surface Range Envelop (SRE)).

truly appropriate dataset and provide a better prediction of assessing the species-environment relationships. Our study has shown that collecting information on the model-based data integration, making a separate sub-model of each dataset, and then integrating them into a single final model could be more accurate than that of data-pooling technique where different data sources were combined and a single model was fitted (Fletcher Jr et al., 2019; Isaac et al., 2020). We examined how various bird data sources can be treated differently in SDMs applying two different techniques of data combination.

Our result also showed that machine-learning algorithm particularly tree-based ensembles Random Forest (RF) (Tonini et al., 2020) is the most accurate modelling technique for predicting bird species distribution and could offer a more precise prediction of assessing the species-environment interactions comparing to other modelling techniques (Li et al., 2017; Mi et al., 2017; Wellmann et al., 2020). This recent algorithm is one of the most accurate techniques in ecological modelling (Bradter et al., 2013; Li and Wang, 2013) that can better model and implement complex non-linear interactions between species and the ecosystem (Garzon et al., 2006; Heikkinen et al., 2012; Oliver et al., 2012). Random Forest is an effective technique for incomplete data and unbalanced databases (Breiman, 2001; Li, 2013) and is reliable to overfit, and commonly produces better predictive models (Howard et al., 2014; Tonini et al., 2020).

5. Conclusions

Our analysis offers robust data-integration approaches to limited, low-distribution structural data, resulting in a more accurate prediction of bird distribution across a large mountain area of the western Swiss Alps. Data-integration has a tremendous ability to better explain the distributions of species and statistical models (Fithian et al., 2015; Fletcher Jr et al., 2019; Miller et al., 2019). For instance, by combining several sources of data in development of a model, a researcher can obtain a better understanding of the environmental relations and mechanisms that affect the species distribution (Fukaya et al., 2020; Isaac et al., 2020; Pacifici et al., 2017). Where there is a lack of data available for a given species, the data integration approach would be able to account for various sampling biases affecting the data and improve the models' predictive performance by increasing the quantity of available data and optimizing the useful information for predicting species distributions (Fithian et al., 2015; Fletcher Jr et al., 2019; Isaac et al., 2020). With rising access to a growing amount of data, an data-integrating approach is expected to offer an increasingly widespread and effective solution to species distribution concerns and emerging issues of environmental change (Miller et al., 2019; Pacifici et al., 2017). We concentrated exclusively on integrating bird distributional data, while other types of data could also be integrated to help predict the species distribution processes more precisely.

Data availability statement

The data that support the findings of this study are available from the Swiss Ornithological Institute but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

Author contributions

NAT designed the study, applied the methodology, and analyzed data with input from all authors. NAT led the writing of the manuscript. All authors contributed critically to the drafts and revised, read, and approved the final manuscript.

Declaration of Competing Interest

None.

Acknowledgements

This work was supported by University of Lausanne and Iran's Ministry of Science, Research and Technology. We would like to thank Dr. Thomas Sattler from the Swiss Ornithological Institute for providing species data and collaboration. We thank Prof. Mikhail Kanevski, Dr. Federico Amato and Fabian Guignard for their valuable comments on an early stage of this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2021.101501.

References

Abadi, F., Gimenez, O., Arlettaz, R., Schaub, M., 2010. An assessment of integrated population models: bias, accuracy, and violation of the assumption of independence. Ecology 91 (1), 7–14.
Aguirre-Gutiérrez, J., Carvalheiro, L.G., Polce, C., van Loon, E.E., Raes, N., Reemer, M., Biesmeijer, J.C., 2013. Fit-for-purpose: species distribution model performance depends on evaluation criteria–Dutch hoverflies as a case study. PLoS One 8 (5), e63708.
Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J. Appl. Ecol. 43 (6), 1223–1232.
Amini Tehrani, N., Naimi, B., Jaboyedoff, M., 2020. Toward community predictions: multi-scale modelling of mountain breeding birds' habitat suitability, landscape preferences, and environmental drivers. Ecol. Evol. 10, 5544–5557. https://doi.org/10.1002/ece3.6295.
Amini Tehrani, N., Naimi, B., Jaboyedoff, M., 2021. Modelling current and future species distribution of breeding birds as regional essential biodiversity variables (SD EBVs): a bird perspective in Swiss Alps. Glob. Ecol. Conserv. https://doi.org/10.1016/j.gecco.2021.e01596.
Araujo, M.B., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. J. Biogeogr. 33 (10), 1677–1688.
Barbet-Massin, M., Jiguet, F., Albert, C.H., Thuiller, W., 2012. Selecting pseudo-absences for species distribution models: how, where and how many? Methods Ecol. Evol. 3 (2), 327–338.
Bikkina, V., 2014. Comparison of Machine Learning Methods for Predicting Bird Distributions. Oregon State University.
Bird, T.J., Bates, A.E., Lefcheck, J.S., Hill, N.A., Thomson, R.J., Edgar, G.J., Pecl, G.T., 2014. Statistical solutions for error and bias in global citizen science datasets. Biol. Conserv. 173, 144–154.
Birrer, S., 2019. Vogelwelt auf dem Golfplatz Andermatt 2019. Schweizerische Vogelwarte, Sempach.
Booth, T.H., Nix, H.A., Busby, J.R., Hutchinson, M.F., 2014. Bioclim: the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. Divers. Distrib. 20, 1–9. https://doi.org/10.1111/ddi.12144.
Bradter, U., Kunin, W.E., Altringham, J.D., Thom, T.J., Benton, T.G., 2013. Identifying appropriate spatial scales of predictors in species distribution models with the random forest algorithm. Methods Ecol. Evol. 4 (2), 167–174.
Braunisch, V., Suchant, R., 2010. Predicting species distributions based on incomplete survey data: the trade-off between precision and scale. Ecography 33 (5), 826–840.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.
Brooks, T.M., Mittermeier, R.A., da Fonseca, G.A., Gerlach, J., Hoffmann, M., Lamoreux, J.F., Rodrigues, A.S., 2006. Global biodiversity conservation priorities. science 313 (5783), 58–61.
Brotons, L., Herrando, S., Pla, M., 2007. Updating bird species distribution at large spatial scales: applications of habitat modelling to data from long-term monitoring programs. Divers. Distrib. 13 (3), 276–288.
Carvalho, S.B., Brito, J.C., Pressey, R.L., Crespo, E., Possingham, H.P., 2010. Simulating the effects of using different types of species distribution data in reserve selection. Biol. Conserv. 143 (2), 426–438.
Carvalho, S.B., Brito, J.C., Crespo, E.G., Watts, M.E., Possingham, H.P., 2011. Conservation planning under climate change: toward accounting for uncertainty in predicted species distributions to increase confidence in conservation investments in space and time. Biol. Conserv. 144 (7), 2020–2030.
Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46.
Descombes, P., Vittoz, P., Guisan, A., Pellissier, L., 2017. Uneven rate of plant turnover along elevation in grasslands. Alp. Bot. 127 (1), 53–63.
Dickinson, J.L., Zuckerberg, B., Bonter, D.N., 2010. Citizen science as an ecological research tool: challenges and benefits. Annu. Rev. Ecol. Evol. Syst. 41, 149–172.
Domisch, S., Wilson, A.M., Jetz, W., 2016. Model-based integration of observed and expert-based information for assessing the geographic and environmental distribution of freshwater species. Ecography 39 (11), 1078–1088.
Dorazio, R.M., 2014. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. Glob. Ecol. Biogeogr. 23 (12), 1472–1484.

Dormann, C.F., Purschke, O., García Márquez, J.R., Lautenbach, S., Schröder, S., 2008. Components of uncertainty in species distribution analysis: a case study of the Great Grey shrike. Ecology 89, 3371–3386.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Münkemüller, T., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36 (1), 27–46.

Edwards Jr., T.C., Cutler, D.R., Zimmermann, N.E., Geiser, L., Moisen, G.G., 2006. Effects of sample survey design on the accuracy of classification tree models in species distribution models. Ecol. Model. 199 (2), 132–141.

Elith, J., Graham, H.C., Anderson, P.R., Dudík, M., Ferrier, S., Guisan, A., Li, J., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29 (2), 129–151.

Fajardo, J., Lessmann, J., Bonaccorso, E., Devenish, C., Munoz, J., 2014. Combined use of systematic conservation planning, species distribution modelling, and connectivity analysis reveals severe conservation gaps in a megadiverse country (Peru). PLoS One 9 (12), e114367.

Fernandes, R.F., Scherrer, D., Guisan, A., 2019. Effects of simulated observation errors on the performance of species distribution models. Divers. Distrib. 25 (3), 400–413.

Ferrier, S., Drielsma, M., Manion, G., Watson, G., 2002. Extended statistical approaches to modelling spatial pattern in biodiversity in Northeast New South Wales. II. Community-level modelling. Biodivers. Conserv. 11 (12), 2309–2338.

Fithian, W., Elith, J., Hastie, T., Keith, D.A., 2014. A proportional observer bias model for multispecies distribution modeling. arXiv preprint. arXiv:1403.7274.

Fithian, W., Elith, J., Hastie, T., Keith, D.A., 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods Ecol. Evol. 6 (4), 424–438.

Fletcher Jr., R.J., Hefley, T.J., Robertson, E.P., Zuckerberg, B., McCleery, R.A., Dorazio, R.M., 2019. A practical guide for combining data to model species distributions. Ecology 100 (6), e02710.

Fletcher, R.J., McCleery, R.A., Greene, D.U., Tye, C.A., 2016. Integrated models that unite local and regional data reveal larger-scale environmental relationships and improve predictions of species distributions. Landsc. Ecol. 31 (6), 1369–1382.

Fukaya, K., Kusumoto, B., Shiono, T., Fujinuma, J., Kubota, Y., 2020. Integrating multiple sources of ecological data to unveil macroscale species abundance. Nat. Commun. 11 (1), 1–14.

Gama, M., Crespo, D., Dolbeth, M., Anastácio, P., 2016. Predicting global habitat suitability for *Corbicula fluminea* using species distribution models: the importance of different environmental datasets. Ecol. Model. 319, 163–169.

Garzon, M.B., Blazek, R., Neteler, M., De Dios, R.S., Ollero, H.S., Furlanello, C., 2006. Predicting habitat suitability with machine learning models: the potential area of *Pinus sylvestris* L. in the Iberian Peninsula. Ecol. Model. 197 (3–4), 383–393.

Ghysels, E., Santa-Clara, P., Valkanov, R., 2004. The MIDAS Touch: Mixed Data Sampling Regression Models. UCLA: Finance. Retrieved from. https://escholarship.org/uc/item/9mf223rs.

Graham, C.H., Elith, J., Hijmans, R.J., Guisan, A., Townsend Peterson, A., Loiselle, B.A., NCEAS Predicting Species Distributions Working Group, 2008. The influence of spatial errors in species occurrence data used in distribution models. J. Appl. Ecol. 45 (1), 239–247.

Grenouillet, G., Buisson, L., Casajus, N., Lek, S., 2011. Ensemble modelling of species distribution: the effects of geographical and environmental ranges. Ecography 34 (1), 9–17.

Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., Wintle, B.A., 2015. Is my species distribution model fit for purpose? Matching data and models to applications. Glob. Ecol. Biogeogr. 24 (3), 276–292.

Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8 (9), 993–1009.

Guisan, A., Thuiller, W., Zimmermann, N.E., 2017. Habitat Suitability and Distribution Models: With Applications in R. Cambridge University Press.

Hallman, T.A., Robinson, W.D., 2020. 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. 26 (3), 315–328.

Heikkinen, R.K., Marmion, M., Luoto, M., 2012. Does the interpolation accuracy of species distribution models come at the expense of transferability? Ecography 35 (3), 276–288.

Hernandez, P.A., Graham, C.H., Master, L.L., Albert, D.L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29 (5), 773–785.

Hirzel, A., Guisan, A., 2002. Which is the optimal sampling strategy for habitat suitability modelling. Ecol. Model. 157 (2–3), 331–341.

Hortal, J., Lobo, J.M., Jiménez-Valverde, A.L.B.E.R.T.O., 2007. Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife, Canary Islands. Conserv. Biol. 21 (3), 853–863.

Howard, C., Stephens, P.A., Pearce-Higgins, J.W., Gregory, R.D., Willis, S.G., 2014. Improving species distribution models: the value of data on abundance. Methods Ecol. Evol. 5 (6), 506–513.

Isaac, N.J., Jarzyna, M.A., Keil, P., Dambly, L.I., Boersch-Supan, P.H., Browning, E., Jarvis, S., 2020. Data integration for large-scale models of species distributions. Trends Ecol. Evol. 35 (1), 56–67.

Jiménez-Valverde, A., 2012. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Glob. Ecol. Biogeogr. 21 (4), 498–507.

Jiménez-Valverde, A., Lobo, J., Hortal, J., 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. Commun. Ecol. 10 (2), 196–205.

Koshkina, V., Wang, Y., Gordon, A., Dorazio, R.M., White, M., Stone, L., 2017. Integrated species distribution models: combining presence-background data and site-occupancy data with imperfect detection. Methods Ecol. Evol. 8 (4), 420–430.

Li, J., 2013, December. Predicting the spatial distribution of seabed gravel content using random forest, spatial interpolation methods and their hybrid methods. In: The International Congress on Modelling and Simulation (MODSIM), pp. 1–6.

Li, J., Alvarez, B., Siwabessy, J., Tran, M., Huang, Z., Przeslawski, R., Nichol, S., 2017. Application of random forest, generalised linear model and their hybrid methods with geostatistical techniques to count data: predicting sponge species richness. Environ. Model. Softw. 97, 112–129.

Li, X., Wang, Y., 2013. Applying various algorithms for species distribution modelling. Integr. Zool. 8 (2), 124–135.

Li, Z., Zhu, Z., Wu, Y., 2020. Scale dependency of pseudo-absences selection and uncertainty in climate scenarios matter when assessing potential distribution of a rare poppy plant Meconopsis punicea maxim. Under a warming climate. Glob. Ecol. Conserv. 24, e01353.

Liu, C., White, M., Newell, G., Griffioen, P., 2013. Species distribution modelling for conservation planning in Victoria, Australia. Ecol. Model. 249, 68–74.

Manel, S., Dias, J.M., Ormerod, S.J., 1999. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. Ecol. Model. 120 (2–3), 337–347.

Marini, M.Â., Barbet-Massin, M., Lopes, L.E., Jiguet, F., 2010. Predicting the occurrence of rare Brazilian birds with species distribution models. J. Ornithol. 151 (4), 857–866.

Marmion, M., Luoto, M., Heikkinen, R.K., Thuiller, W., 2009. The performance of state-of-the-art modelling techniques depends on geographical distribution of species. Ecol. Model. 220 (24), 3512–3520.

Mendes, P., Velazco, S.J.E., de Andrade, A.F.A., Júnior, P.D.M., 2020. Dealing with overprediction in species distribution models: how adding distance constraints can improve model accuracy. Ecol. Model. 431, 109180.

Merow, C., Wilson, A.M., Jetz, W., 2017. Integrating occurrence data and expert maps for improved species range predictions. Glob. Ecol. Biogeogr. 26 (2), 243–258.

Mi, C., Huettmann, F., Guo, Y., Han, X., Wen, L., 2017. Why choose random forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. PeerJ 5, e2849.

Miller, D.A., Pacifici, K., Sanderlin, J.S., Reich, B.J., 2019. The recent past and promising future for data integration methods to estimate species' distributions. Methods Ecol. Evol. 10 (1), 22–37.

Moudrý, V., Šímová, P., 2012. Influence of positional accuracy, sample size and scale on modelling species distributions: a review. Int. J. Geogr. Inf. Sci. 26 (11), 2083–2095.

Niel, C., Lebreton, J.D., 2005. Using demographic invariants to detect overharvested bird populations from incomplete data. Conserv. Biol. 19 (3), 826–835.

Oliver, T.H., Gillings, S., Girardello, M., Rapacciuolo, G., Brereton, T.M., Siriwardena, G. M., Fuller, R.J., 2012. Population density but not stability can be predicted from species distribution models. J. Appl. Ecol. 49 (3), 581–590.

Pacifici, K., Reich, B.J., Miller, D.A., Gardner, B., Stauffer, G., Singh, S., Collazo, J.A., 2017. Integrating multiple data sources in species distribution modeling: a framework for data fusion. Ecology 98 (3), 840–850.

Parra, J.L., Graham, C.C., Freile, J.F., 2004. Evaluating alternative data sets for ecological niche models of birds in the Andes. Ecography 27 (3), 350–360.

Parviainen, M., Marmion, M., Luoto, M., Thuiller, W., Heikkinen, R.K., 2009. Using summed individual species models and state-of-the-art modelling techniques to identify threatened plant species hotspots. Biol. Conserv. 142 (11), 2501–2509.

Pearson, R.G., Thuiller, W., Araújo, M.B., Martinez-Meyer, E., Brotons, L., McClean, C., Lees, D.C., 2006. Model-based uncertainty in species range prediction. J. Biogeogr. 33 (10), 1704–1711.

Picard, N., Chagneau, P., Mortier, F., Bar-Hen, A., 2009. Finding confidence limits on population growth rates: bootstrap and analytic methods. Math. Biosci. 219 (1), 23–31.

Pressey, R.L., 2004. Conservation planning and biodiversity: assembling the best data for the job. Conserv. Biol. 18 (6), 1677–1681.

R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Randin, C.F., Jaccard, H., Vittoz, P., Yoccoz, N.G., Guisan, A., 2009. Land use improves spatial predictions of mountain plant abundance but not presence-absence. J. Veg. Sci. 20 (6), 996–1008.

Ratnieks, F.L., Schrell, F., Sheppard, R.C., Brown, E., Bristow, O.E., Garbuzov, M., 2016. Data reliability in citizen science: learning curve and the effects of training method, volunteer background and experience on identification accuracy of insects visiting ivy flowers. Methods Ecol. Evol. 7 (10), 1226–1235.

Reiss, H., Cunze, S., König, K., Neumann, H., Kröncke, I., 2011. Species distribution modelling of marine benthos: a North Sea case study. Mar. Ecol. Prog. Ser. 442, 71–86.

Rodríguez, J.P., Brotons, L., Bustamante, J., Seoane, J., 2007. The application of predictive modelling of species distribution to biodiversity conservation. Divers. Distrib. 13 (3), 243–251.

Saracco, J.F., Desante, D.F., Nott, M.P., Hochachka, W.M., Kelling, S., Fink, D., 2008, February. Integrated bird monitoring and the avian knowledge network: Using multiple data resources to understand spatiotemporal variation in demographic processes and abundance. In: Tundra to Tropics: Connecting Birds, Habitats and People. Proceedings of the 4th International Partners in Flight Conference, pp. 13–16.

Schank, C.J., Cove, M.V., Kelly, M.J., Nielsen, C.K., O'Farrill, G., Meyer, N., Dobbins, M., 2019. A sensitivity analysis of the application of integrated species distribution models to Mobile species: a case study with the endangered Baird's tapir. Environ. Conserv. 46 (3), 184–192.

Scherrer, D., Christe, P., Guisan, A., 2019. Modelling bat distributions and diversity in a mountain landscape using focal predictors in ensemble of small models. Divers. Distrib. 25 (5), 770–782.

Schmid, H., Zbinden, N., Keller, V., 2004. Überwachung der Bestandsentwicklung Häufiger Brutvögel in der Schweiz. Swiss Ornithological Institute Sempach Switzerland.

Segurado, P., Araujo, M.B., 2004. An evaluation of methods for modelling species distributions. J. Biogeogr. 31 (10), 1555–1568.

Simmonds, E.G., Jarvis, S.G., Henrys, P.A., Isaac, N.J., O'Hara, R.B., 2020. Is more data always better? A simulation study of benefits and limitations of integrated distribution models. Ecography 43 (10), 1413–1422.

Smeraldo, S., Bosso, L., Fraissinet, M., Bordignon, L., Brunelli, M., Ancillotto, L., Russo, D., 2020. Modelling risks posed by wind turbines and power lines to soaring birds: the black stork (*Ciconia nigra*) in Italy as a case study. Biodivers. Conserv. 1–18.

Smeraldo, S., Bosso, L., Salinas-Ramos, V.B., Ancillotto, L., Sánchez-Cordero, V., Gazaryan, S., Russo, D., 2021. Generalists yet different: distributional responses to climate change may vary in opportunistic bat species sharing similar ecological traits. Mammal Rev.

Steen, V.A., Elphick, C.S., Tingley, M.W., 2019. An evaluation of stringent filtering to improve species distribution models from citizen science data. Divers. Distrib. 25 (12), 1857–1869.

Talluto, M.V., Boulangeat, I., Ameztegui, A., Aubin, I., Berteaux, D., Butler, A., Liénard, J., 2016. Cross-scale integration of knowledge for predicting species ranges: a metamodelling framework. Glob. Ecol. Biogeogr. 25 (2), 238–249.

Tenan, S., Pedrini, P., Bragalanti, N., Groff, C., Sutherland, C., 2017. Data integration for inference about spatial processes: A model-based approach to test and account for data inconsistency. PLoS One 12 (10), e0185588.

Thuiller, W., 2003. BIOMOD–optimizing predictions of species distributions and projecting potential future shifts under global change. Glob. Chang. Biol. 9 (10), 1353–1362.

Thuiller, W., Georges, D., Engler, R., Breiner, F., Georges, M.D., Thuiller, C.W., 2016. Package 'biomod2`. Species Distribution Modeling within an Ensemble Forecasting Framework. Software.

Tonini, M., D'Andrea, M., Biondi, G., Degli Esposti, S., Trucchia, A., Fiorucci, P., 2020. A machine learning-based approach for wildfire susceptibility mapping. The case study of the Liguria region in Italy. Geosciences 10 (3), 105.

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D., Kadmon, R., 2007. A comparative evaluation of presence-only methods for modelling species distribution. Divers. Distrib. 13 (4), 397–405.

Tulloch, A.I., Sutcliffe, P., Naujokaitis-Lewis, I., Tingley, R., Brotons, L., Ferraz, K.M.P., Rhodes, J.R., 2016. Conservation planners tend to ignore improved accuracy of modelled species distributions to focus on multiple threats and ecological processes. Biol. Conserv. 199, 157–171.

Wagner, T., Hansen, G.J., Schliep, E.M., Bethke, B.J., Honsey, A.E., Jacobson, P.C., White, S.L., 2020. Improved understanding and prediction of freshwater fish communities through the use of joint species distribution models. Can. J. Fish. Aquat. Sci. 77 (9), 1540–1551.

Watling, J.I., Brandt, L.A., Bucklin, D.N., Fujisaki, I., Mazzotti, F.J., Romanach, S.S., Speroterra, C., 2015. Performance metrics and variance partitioning reveal sources of uncertainty in species distribution models. Ecol. Model. 309, 48–59.

Wellmann, T., Lausch, A., Scheuer, S., Haase, D., 2020. Earth observation based indication for avian species distribution models using the spectral trait concept and machine learning in an urban setting. Ecol. Indic. 111, 106029.

Wilson, K.A., Westphal, M.I., Possingham, H.P., Elith, J., 2005. Sensitivity of conservation planning to different approaches to using predicted species distribution data. Biol. Conserv. 122 (1), 99–112.

Wilson, R.J., Thomas, C.D., Fox, R., Roy, D.B., Kunin, W.E., 2004. Spatial patterns in species distributions reveal biodiversity change. Nature 432 (7015), 393–396.

Zhang, X., Vincent, A.C., 2017. Integrating multiple datasets with species distribution models to inform conservation of the poorly-recorded Chinese seahorses. Biol. Conserv. 211, 161–171.