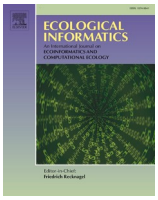
Ecological Informatics 72 (2022) 101914

Contents lists available at ScienceDirect

Ecological Informatics

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



Background sampling for multi-scale ensemble habitat selection modeling: Does the number of points matter?

Logan Hysen [a,*], Danial Nayeri [a], Samuel Cushman [b], Ho Yi Wan [a]

[a] Department of Wildlife, California State Polytechnic University Humboldt, 1 Harpst Street, Arcata 95521, CA, USA
[b] USDA, Rocky Mountain Research Station, Flagstaff, AZ, USA

ARTICLE INFO

Keywords:
Biomod
Ecological niche
Habitat suitability
Presence-absence
Species distribution model
Presence-only

ABSTRACT

Ensemble habitat selection modeling is becoming a popular approach among ecologists to answer different questions. Since we are still in the early stages of development and application of ensemble modeling, there remain many questions regarding performance and parameterization. One important gap, which this paper addresses, is how the number of background points used to train models influences the performance of the ensemble model. We used an empirical presence-only dataset and three different selections of background points to train scale-optimized habitat selection models using six modeling algorithms (GLM, GAM, MARS, ANN, Random Forest, and MaxEnt). We tested four ensemble models using different combinations of the component models: (a) equal numbers of background points and presences, (b) background points equaled ten times the number of presences, (c) 10,000 background points, and (d) optimized background points for each component model. Among regression-based approaches, MARS performed best when built with 10,000 background points. Among machine learning models, RF performed the best when built with equal presences and background points. Among the four ensemble models, AUC indicated that the best performing model was the ensemble with each component model including the optimized number of background points, while TSS increased as the number of background points models increased. We found that an ensemble of models, each trained with an optimal number of background points, outperformed ensembles of models trained with the same number of background points, although differences in performance were slight. When using a single modeling method, RF with equal number of presences and background points can perform better than an ensemble model, but the performance fluctuates when the number of background points is not properly selected. On the other hand, ensemble modeling provides consistently high accuracy regardless of background point sampling approach. Further, optimizing the number of background points for each component model within an ensemble model can provide the best model improvement. We suggest evaluating more models across multiple species to investigate how background point selection might affect ensemble models in different scenarios.

1. Introduction

Habitat selection models are widely used, popular tools for understanding the spatial patterns of habitat or distribution for a species or group of species (Elith* et al., 2006; Guisan and Zimmermann, 2000; Hegel et al., 2010). There are many approaches to building a habitat selection model, each with their own unique strengths and weaknesses (Valavi et al., 2021). To reconcile the differences among these approaches, researchers use a modeling procedure called ensemble modeling, which combines the predictions from multiple modeling approaches that each reflect a possible state of the system in question into a consensus (Araújo et al., 2005; Araújo and New, 2007; Torres et al., 2010). In practice and to make predictions, the consensus is an ensemble model that reflects the central tendency of the component models (Araújo et al., 2005; Thuiller, 2004; Zhu et al., 2021), which can be a combination of both regression-based and machine learning algorithms (Araújo and New, 2007; Mohammadi et al., 2022; Thuiller, 2004). To build the consensus, researchers can average model coefficients before making predictions or take the average of model predictions (Araújo and New, 2007; Burnham and Anderson, 2002; Gregory et al., 2001). Either way, models can be weighted equally or weighted differently by how well they perform, assigning higher weights to better-performing models

* Corresponding author.
E-mail address: lbh22@humboldt.edu (L. Hysen).

https://doi.org/10.1016/j.ecoinf.2022.101914
Received 19 August 2022; Received in revised form 18 October 2022; Accepted 9 November 2022
Available online 12 November 2022

(Thuiller, 2004; Zhu et al., 2021). The advantage of combining different types of models is that it combines the strengths of all models to compensate for the weaknesses of any particular model, strengthening the robustness of the final product (Araújo and New, 2007; Bates and Granger, 1969). Increasingly, studies have demonstrated the superiority of ensemble modeling over individual modeling approaches in predicting species-environment relationships (Araújo and New, 2007; Marmion et al., 2009; Mohammadi et al., 2022; but see Hao et al., 2020).

As with any individual habitat selection model, training the component models for ensemble modeling requires location information, or presence-absence data, that represents whether the study species was detected or not at the sampling site (Barbet-Massin et al., 2012; Elith and Leathwick, 2009). Because most surveying and monitoring efforts are designed to detect the occurrences and not the non-occurrence of a species, absence data is often unavailable (VanDerWal et al., 2009). Therefore, artificially created background points, or pseudo-absences, are often used as substitutes to represent the environmental conditions available to a species (VanDerWal et al., 2009). Previous studies have assessed the optimal number of background points for individual modeling approaches (Barbet-Massin et al., 2012; Liu et al., 2019; VanDerWal et al., 2009). In general, regression-based approaches like generalized linear models (GLMs) and generalized additive models (GAMs) perform best when the number of background points substantially exceeds the number of presence points, whereas machine learning algorithms like Random Forest (RF) and Artificial Neural Networks (ANN) typically perform best when the number of background points and presences are roughly equal (Barbet-Massin et al., 2012, but see Cushman et al., 2017). For RF and logistic regression in particular, care must be taken to avoid bias and the selection of background points should be representative of the heterogeneity of the landscape (Cushman et al., 2017). Other exceptions exist, as well. For example, multivariate adaptive regression splines (MARS), a regression-based approach, has been shown to perform best when the number of presences and background points are roughly equal (Barbet-Massin et al., 2012). Furthermore, the popular software MaxEnt, a machine learning approach, is known to perform well with a large number of background points (Barbet-Massin et al., 2012). Although the information is useful, the within-group inconsistencies described above can be confusing when modelers need to decide the number of background points to use in the context of ensemble modeling. It remains unclear how the number of background points selected to train each component model might influence the overall performance of an ensemble model.

Species-environment relationships are often scale-dependent and as a result it is important to assess the spatial scale at which an ecological process or species-environment relationship occurs (Wiens, 1989). This knowledge has resulted in the development of multi-scale habitat selection models, which identify the spatial scale at which different environmental variables are important for habitat selection by a species, and then combine those variables at their optimal spatial scales into one final model (McGarigal et al., 2016). Combining multi-scale modeling with ensemble modeling has the potential to harness the benefits of both approaches, but because we are still in the early stages of applying them together, it is unknown whether and how the number of background points might affect this approach.

To unlock the full potential of multi-scale ensemble modeling in predicting habitat suitability and species occurrence, we must investigate the sensitivity of ensemble model performance to the number of background points. To test this, we conducted multi-scale ensemble modeling in an empirical setting using three regression-based methods (i.e., GLM, GAM, MARS) and three machine-learning methods (i.e., MaxEnt, RF, ANN), and assessed the performance of four different scenarios of background point selection: (a) one where all component models were built with the same number of background points as presences, (b) one where the number of background points equaled ten times the number of presences, (c) one with 10,000 background.

points, and (d) one where each component model was built with an optimal number of background points.

2. Materials and methods

2.1. Study area

The study area is located in the Redwood Coast ecoregion of northwestern California (Fig. 1; Ecoregion Level III 263a; United States Forest Service, 2017). This region encompasses 16,500 km2 and is dominated by redwoods and tall evergreen trees with patches of broadleaf woodlands in addition to coastal scrubland (USFS, 2017). The Redwood Coast ecoregion is part of the California Floristic Province and is home to multiple endemic species, making the ecosystems in the region some of the most unique in North America (Conservation International, 2022).

2.2. Species data

To achieve results with real-world applicability, we obtained 248 nesting locations from 2015 to 2020 for the threatened northern spotted owl (*Strix occidentalis caurina*) from the California Department of Fish and Wildlife. Northern spotted owls often use the same home range for nesting each year and are typically monogamous, although they do not always use the same nesting structure each year, meaning that multiple nesting locations could be recorded for each owl pair across different years. To avoid pseudo replication of owl pairs, we filtered the nesting locations using unique owl identification numbers so that only the most recently recorded nesting location for a given pair of northern spotted owl was used to train and test models.

2.3. Environmental data

For the modeling, we selected 24 a priori predictor covariates related to northern spotted owl habitat selection based on our review of the scientific literature. These covariates cover the categories of

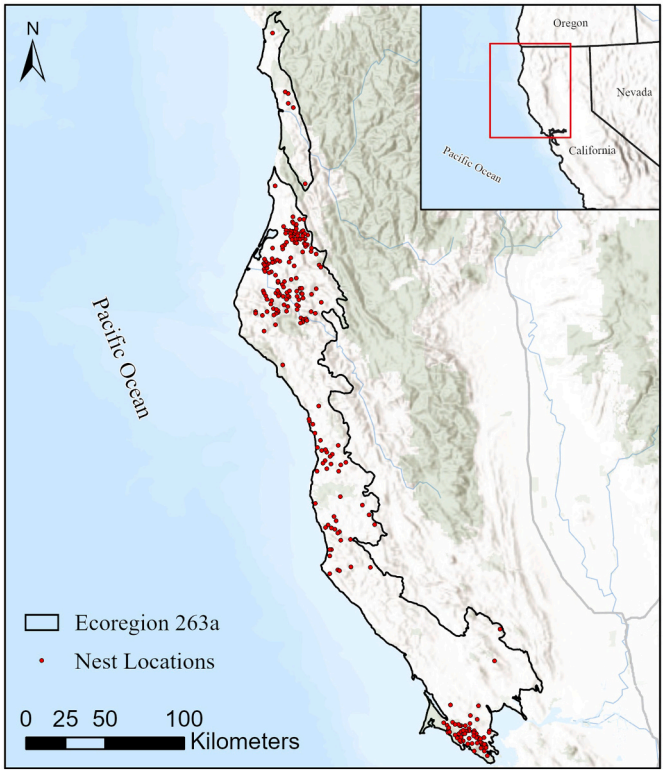


Fig. 1. Showing the study area and northern spotted owl nesting locations (n = 248).

composition (e.g., forest structure), topography, and climate (Table S1). The covariates were obtained from the Landscape Ecology, Modeling, Mapping, and Analysis Gradient Nearest Neighbor (LEMMA GNN) database from Oregon State University (Ohmann and Gregory, 2002), Landscape Fire and Resource Management Planning Tools Project (LANDFIRE, 2017), and the PRISM climate group at Oregon State University (Rupp et al., 2022). Curvature, slope, and insolation (i.e., solar radiation index) were calculated from a digital elevation model obtained from LANDFIRE using Spatial Analyst Tools in ArcGIS 10.3. All covariates were in raster format, projected into the NAD 1983 / UTM 11 N projection, and resampled to 30 m resolution if necessary.

2.4. Background point generation

To identify the region in which to generate background points, we subtracted 1.3 km buffers around each nesting location from a buffer of 30 km around all nesting locations. We chose the upper limit (30-km from nesting locations) based on the median dispersal ability for northern spotted owls to limit artificial inflation of test statistics (Forsman et al., 2002; United States Fish and Wildlife Service, 2011), and the lower limit (1.3-km from nesting locations) based on average nesting season home range sizes in the region to avoid generating background points where we know there are northern spotted owls using habitat (VanDerWal et al., 2009; Weisel, 2015). We generated 20,000 background locations using a random approach, which has been shown to be an acceptable method for background point generation (Barbet-Massin et al., 2012). We then subsampled background locations from this total to ensure models were built using the same available data.

2.5. Spatial scaling

Since the relationship between a species and its environment varies with spatial scale (McGarigal et al., 2016), we sought to identify the optimal scale of effect for each covariate in our models. To do this, we conducted univariate scaling across a range of biologically meaningful scales for each covariate. To do this, we conducted focal statistics on each covariate using a circular moving window radius starting at 100 m, then increasing in a geometric doubling sequence for five iterations (i.e., 200 m, 400 m, 800 m, 1600 m, and 3200 m radius). This produced a set of six rasters for each covariate, each representing a different scale. Then, we extracted values from each scale raster to our nest locations and background locations. To compare values extracted at nesting locations to background locations, we used *t*-tests for each scale and selected the optimal scale as the one that showed a greater difference between the mean of values extracted at nesting locations and the mean of values extracted at background locations. We repeated this procedure for each covariate. To protect against multicollinearity in our models, we compared each covariate at their optimal scales using Pearson's correlation coefficient. For any pair of covariates that had a correlation of $r > |0.7|$, we discarded the covariate that had a smaller difference between the mean of values extracted at nesting locations and the mean of values extracted at background locations.

To reduce spatial autocorrelation in regression-based models and avoid thinning presence points, we built a spatial autocovariate and included it in the GLMs, GAMs, and MARS. We calculated this spatial autocovariate by using the distance weighted average of neighboring response values, given by Eq. 1 in the supplementary material (Dormann et al., 2007).

2.6. Evaluation and training datasets

To ensure that all models were evaluated with the same relatively independent data, we used a common approach to split the data into evaluation and training datasets prior to training the models. We set aside the evaluation dataset and used it to evaluate individual averaged modeling approaches and the final ensemble models, hereafter referred to as external evaluation. Crucially, this data was not used during the model training process and was just used for evaluation purposes. To train and cross-validate (hereafter referred to as internal evaluation) replicates of component models we used the remaining data.

To separate data into training and testing datasets, we used a spatial blocking technique, in which we divided our study area into 25 equally sized blocks, each approximately 41 km × 41 km in size, using the blockCV package in the R programming language (Valavi et al., 2019). We determined the size of our spatial blocks using the spatialAutoRange function in the blockCV package, which returns a block size equal to the median range of spatial autocorrelation in independent covariates (Valavi et al., 2019). This attempts to ensure data within a block is relatively independent of data outside of that block. We then divided these blocks among five folds in which the numbers of presences and background points were roughly equal. We used five folds because, due to the spatial arrangement of our presence data, using more would cause folds to have drastically differing numbers of presences and background points. Then, to obtain presences and background points to use during external evaluation, we set aside one fold. We then used the remaining four folds for internal cross-validation during model training.

2.7. Subsetting background points

We made three random selections of background points from the training dataset for use during model construction. First, we made the "1×" selection, which selected the same number of background points as presences in the training dataset. Second, we made the "10×" selection, which selected a number of background points equal to ten times the number of presences in the training dataset. Third, we made the "10 K" selection, which selected 10,000 background points from the training dataset. We used this approach to maintain consistency across background points used to train each model so there was no variation in model results due to variation in the response variable.

2.8. Training individual models

To train the component models for the ultimate ensemble models, we used six modeling approaches, categorized into regression-based and machine learning models. Regression-based models included Generalized Linear Models (GLMs), Generalized Additive Models (GAMs), and Multivariate Adaptive Regression Splines (MARS). Machine learning models included Random Forest (RF), Artificial Neural Networks (ANN), and Maximum Entropy software (MaxEnt). We built all models using the biomod2 package in R (Thuiller et al., 2021).

To test the influence of the number of selected background points on each modeling approach, we individually applied each model three times, each time using a different background point selection (Fig. 2). For each background point selection, we ran each modeling approach four times, using three of the folds for training and the fourth for internal evaluation, rotating them each replicate so each fold was used for evaluation once. We then averaged the four replicates for each model. Therefore, we obtained three averaged models for each modeling approach, one for each background point selection, for a total of 18 models (Fig. 2). Each model was trained with all environmental covariates that were kept. We used default settings for each modeling approach except for MaxEnt and ANN, which were tuned to limit overfitting and allow model convergence, respectively.

2.9. Evaluating individual modeling approaches

To assess the optimal number of background points for each individual modeling approach, we calculated the True Skill Statistic (TSS) and the Area Under the Receiver Operating Curve (AUC-ROC) for each model using the external evaluation datasets set aside prior to training models (Allouche et al., 2006; Hao et al., 2020; Shabani et al., 2016). The True Skill Statistic is based on a binary classification of model

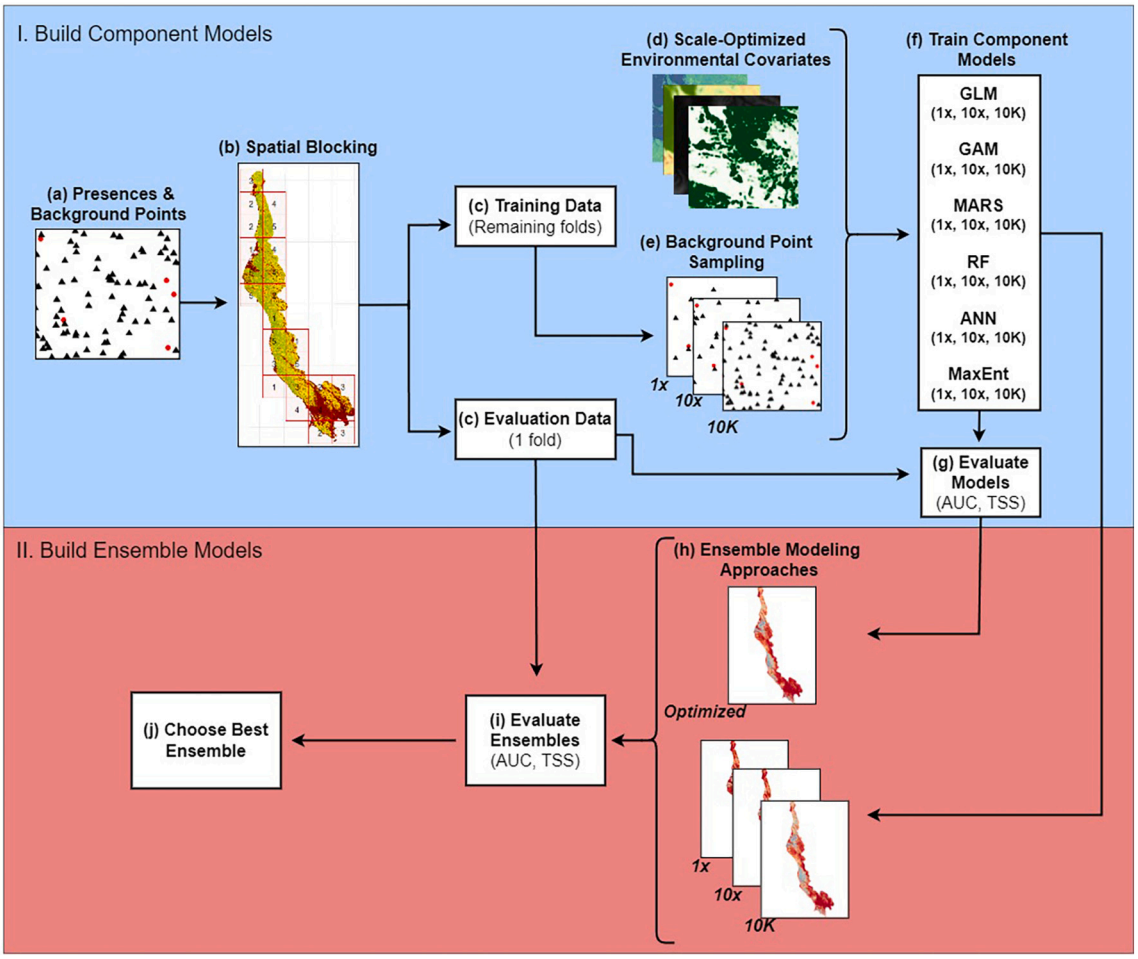


Fig. 2. Diagram showing the procedure used to build models. (a) Obtain presences (red circles) and generate background points (black triangles). (b) Separate data into folds. (c) Set aside one fold to use as an external evaluation dataset and keep the remaining folds as training data. (d) Perform scale optimization for environmental covariates: calculate focal statistics for each covariate using a moving window of radius equal to each scale to be tested, then compare scales using *t*-tests. (e) Subsample background points in the training data to create three datasets: one with the number of background points equal to the number of presences (1×), one with a number of background points equal to 10 times the number of presences (10×), and 10,000 background points (10*K*). Keep the number of presences the same. (f) Train component models separately using each background point subsample. (g) Evaluate component models using the evaluation dataset (using AUC and TSS). (h) Ensemble models using an averaging approach, each time using all six component models. For the optimized ensemble model, for each component model include the one trained with the number of background points with the highest AUC. For the 1×, 10×, and 10 K ensemble models, include all component models built with the 1×, 10×, and 10 K background points, respectively. (i) Evaluate ensemble models using the evaluation dataset (using AUC and TSS here). (j) Compare ensemble model AUCs, TSSs, sensitivities, and specificities and choose the model that performs the best for the data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

predictions into predicted presences and predicted absences (traditionally), meaning that a threshold must be chosen to determine at what value a prediction is considered a "presence" versus an "absence." We selected the threshold for each model by maximizing the sum of sensitivity and specificity (maxSSS) (Liu et al., 2013; Liu et al., 2016). The AUC is a threshold-independent method of total discrimination between observed presence vs background across the range of predicted occurrence probability. We then considered the three models built for each modeling approach and compared model evaluation metrics to choose the model that appeared to represent the data best. We considered the number of background points used to train that best model as the optimal number of background points for that modeling approach for our data.

2.10. Building ensemble models

To determine the influence of the number of background points used to train individual models on the final ensemble model, we ensembled models using an averaging approach, in which each model was weighted equally, for four different combinations of individual models: (1) using allsix individual modeling approaches built with an equal number of background points as presences (i.e., 1×), (2) using all six individual modeling approaches built with the number of background points equal to ten times the number of presences (i.e., 10×), (3) using all six individual modeling approaches built with 10,000 background points (i.e., 10 K), and (4) using all six individual modeling approaches built with their optimal number of background points, which we refer to as the optimized ensemble model (i.e., 1×, 10×, or 10 K).

2.11. Model evaluation and comparison

For each ensemble model, we calculated the TSS, AUC, specificity, and sensitivity using the evaluation dataset set aside prior to training models. To determine the effect of different numbers of background points on the ensemble models, we compared these evaluation metrics between models. The entire procedure is summarized in Fig. 2.

3. Results

The final filtered dataset contained the most recent nesting location for 248 unique pairs of breeding northern spotted owls. After splitting the dataset into evaluation and training data, 72 nesting locations were used for evaluation and 176 nesting locations were used for training.

Then, the background points in the training dataset were subsetted to create the three background point selections (1× = 176, 10× = 1760, 10 K = 10,000). In addition, after assessing correlation between explanatory variables, 12 were removed and 12 were retained (not including the spatial autocovariate) (Table S1).

3.1. Individual modeling approaches

We built 18 models using all environmental variables and six modeling algorithms across three different background point selections using a fixed number of presences and compared the performance of models using AUC and TSS (Fig. 3, Table S2). Among regression-based models, MARS performed the best with an AUC of 0.819 and a TSS of 0.470 (threshold = 0.315) when built with the 10 K background point selection. Among machine learning models, RF performed the best with an AUC of 0.855 when built with the 1× background point selection. However, when comparing models using TSS, MaxEnt performed the best with a TSS of 0.533 (threshold = 0.180) when built with the 10× background point selection. The worst performing model was a GAM with an AUC of 0.647 when built with the 10× background points selection. However, when comparing models using TSS, RF performed the worst when built with the 10 K background point selection, with a TSS of 0.213 (threshold = 0.260). For each modeling algorithm, we compared AUCs between the three background point selections and selected the model with the highest AUC as the "optimal" number of background points for that particular algorithm for this data, reported in Table S2.

3.2. Ensemble modeling

We built four ensemble models, each built using an averaging approach with a different combination of component models. For the optimized ensemble model, we included the best performing models based on AUC: GLM (10×), GAM (1×), MARS (10 K), MaxEnt (10×), RF (1×), and ANN (1×). Of the four ensemble models, the best performing model was the optimized ensemble with an AUC value of 0.835, which was an AUC 0.024 higher than the worst- performing model, the Ensemble 10 K (Table S3, Fig. 4). However, when comparing models using TSS, the 10 K ensemble performed best with a TSS of 0.492 (threshold = 0.395), with the optimized ensemble performing slightly worse with a TSS of 0.472 (threshold = 0.330). When comparing models using TSS, the worst performing ensemble model was the 1× ensemble with a TSS of 0.452 (threshold = 0.340).

4. Discussion

Ensemble modeling methods are widely used to build habitat selection models, which are often used by managers and conservationists across the world (Coetzee et al., 2009; Ramirez-Reyes et al., 2021; Rew et al., 2020; Zeller et al., 2021). Often, these models are built using the same number of background points for each of the component models. However, for individual modeling techniques, the optimal number of background points used to train those models varies widely based on the approach being used (Barbet-Massin et al., 2012; Liu et al., 2019). However, it remains unclear whether this holds true once these models are ensembled.

4.1. Individual models

Our results support previous findings that it is important to consider the number of background points when training individual habitat

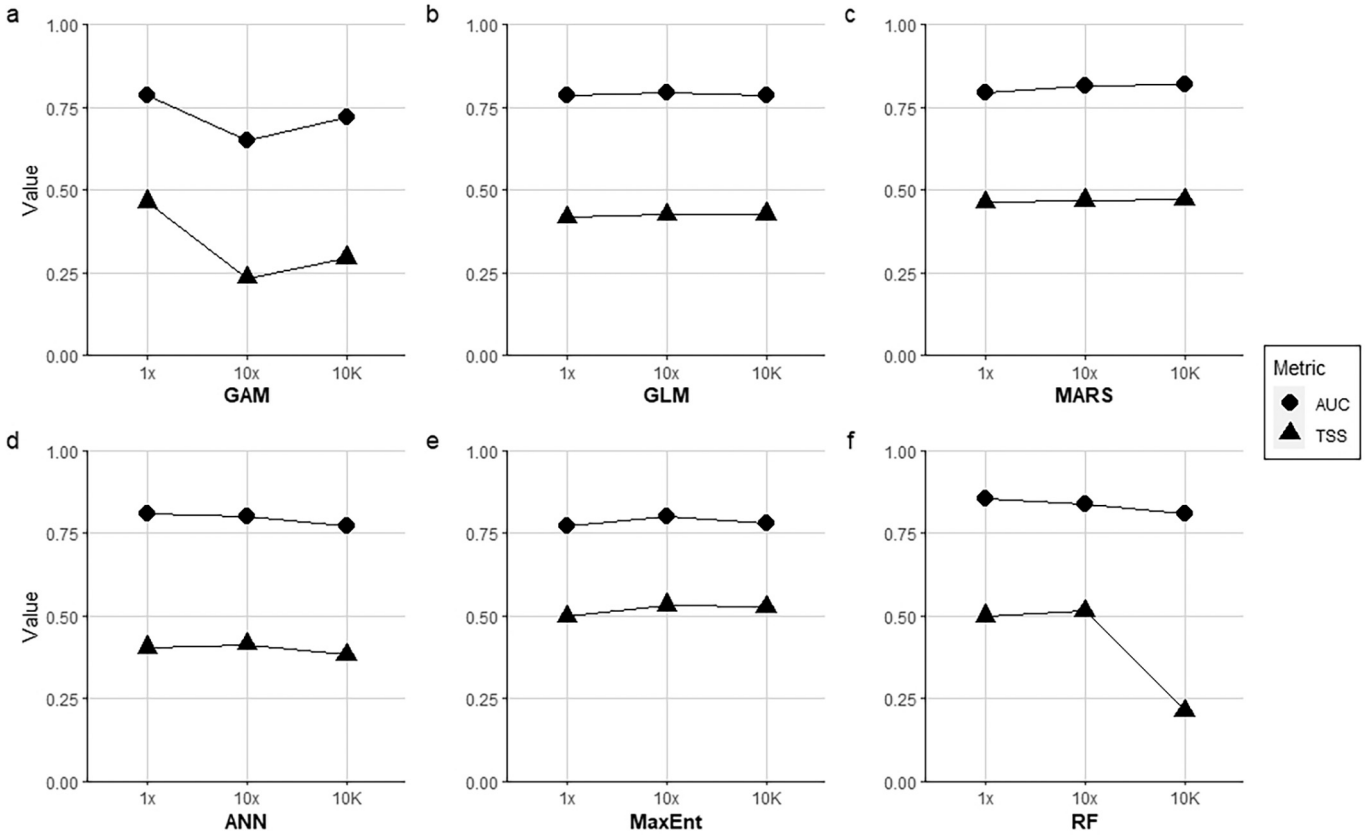


Fig. 3. Performance metrics (i.e., AUC, TSS) for each modeling algorithm across three background point selections (1×, 10×, 10 K).

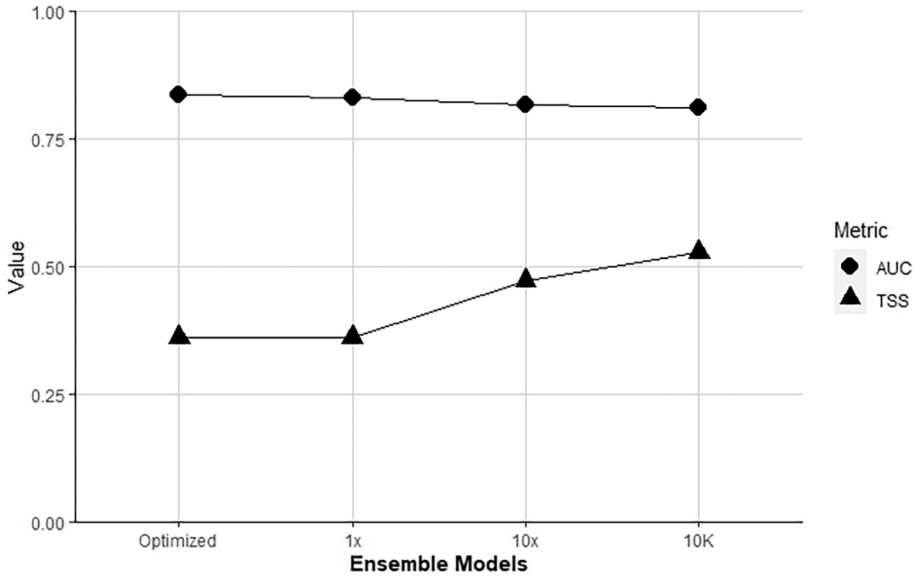


Fig. 4. Performance metrics (i.e., AUC, TSS) for each ensemble model.

selection models (Barbet-Massin et al., 2012). Theoretically, there are infinite background point sampling scenarios that can be tested. However, testing a large number of background point sampling schemes is impractical due to computational time. Therefore, we considered three different background points sampling strategies that are commonly used in research. We chose to use the 1× and 10× selections based on a previous study (Barbet-Massin et al., 2012), and chose the 10 K selection because that is the default setting for the popular modeling software MaxEnt (Phillips et al., 2006). Thus, our findings should be used as a general guideline on the background points sampling strategy rather than recommendations of a specific number. Training models with different numbers of background points impacted the performance metrics, although they did not have the same effect across models or even between the two metrics we compared. For example, the AUC for RF slightly increased as the number of background points used to train the model decreased, while the TSS showed a greater increase (Fig. 3). For the regression- based models, the general trend is in agreement with previous findings, however our findings contradict Barbet-Massin et al. (2012) on individual modeling approaches.

Overall, we found that regression based models tended to perform better with more background points than presences. The GAM was the exception, performing best (based on AUC) with an equal number of background points and presences, contrary to previous findings (Barbet-Massin et al., 2012; Liu et al., 2019). Further, Barbet-Massin et al. (2012) found that MARS performed best with an equal number of background points and presences, but we found support for the opposite relationship. For the machine learning models, our results were in agreement with previous studies (Barbet-Massin et al., 2012; Liu et al., 2019).

We showed that both ANN and RF, when assessed using AUC, performed best with a smaller number of background points, while RF outperformed ANN. However, when assessed with TSS, both ANN and RF performed best with a larger number of background points, although RF still outperformed ANN. In addition, MaxEnt performed best with a larger number of background points. Overall, ANN, MaxEnt, GLM and MARS showed similar performances across all three scenarios, suggesting that these models are less sensitive to changes in background points. Conversely, RF and GAM models show larger discrepancies in performance, suggesting that they might be more prone to background sampling biases.

The discrepancy between results for the regression-based models could signify that the optimal number of background points for a particular model varies between species or even the particular data available. Barbet-Massin et al. (2012) and Liu et al. (2019) used virtual species for their work, while we used empirical presence-only data that is likely more biased than simulated data. However, we argue that it is necessary to test these relationships with both simulated and empirical data, since as we showed, some recommendations may not transfer from the virtual world to the real world. It is possible that more concrete recommendations for the optimal number of background points to use in a given modeling approach will need to wait until a more comprehensive, inter-specific study takes place. We recommend that future studies investigate this issue further across multiple species using publicly available presence-only data. In addition, because we tested only three of the most common background points sampling strategies, we recommend future studies to look at other background points sampling strategies that are perhaps less common.

4.2. Ensemble models

We have provided further evidence that individual modeling approaches perform better when the number of background points used to train each model is optimized for that particular modeling algorithm. Given the popularity of ensemble modeling approaches in the scientific literature, we went further and demonstrated how the number of background points used to train individual models can influence the performance of ensemble models. We tested ensemble models built with four combinations of background points and found that the ensemble performed slightly better (based on AUC) when all individual models used to train the ensemble were built with their optimal number of background points (Fig. 4). Although there is only a small increase in performance, this potentially highlights the ability of ensemble models to balance the strengths and weaknesses of the component models, even when they are not trained with the optimal number of background points (Araújo and New, 2007; Shabani et al., 2016).

Meanwhile, the opposite trend is true when models are assessed using TSS with a threshold chosen by maximizing the sum of sensitivity and specificity. The potential reason for this behavior is clear when assessed in tandem with sensitivity and specificity. When the number of background points used to train all the component models is high, the sensitivity of the model decreases while the specificity increases. When more background points are used during training, the model can become adept at predicting the background at the expense of the presences. In cases using presence-only data, the ability to correctly predict the background is not usually very useful, although it is useful when

answering certain questions (Sofaer et al., 2019). Either way, this provides a clear argument for the need to carefully consider the question at hand and the makeup of data used to train component models prior to performing the analysis. However, the result strongly suggests the need for further work using controlled simulations (e.g., Chiaverini et al., 2021; Atzeni et al., 2020; Kumar et al., 2021) that can generate known habitat relationships. A simulation experiment would then enable objective assessment of which of the metrics (TSS or AUC) is performing correctly, or if both are biased in different ways across the parameter space.

4.3. Consensus-building methods

In addition to considering the number of background points used to train component models in an ensemble modeling framework, other considerations emerge when attempting to find the consensus between models. One such consideration is the method used to weight component models in the ensemble. Some of the most common methods include taking the mean, weight-averaging (WA), and median Principal Components Analysis (mPCA) (Thuiller, 2004; Zhu et al., 2021). To make comparisons between ensemble models simple for this paper and to keep the focus clear, we used the mean approach, which involved averaging each component model using equal weights. As a result, we were able to succinctly demonstrate how the number of background points used to train component models can impact the final ensemble model. However, it is often necessary to assign weights to different component models during consensus-building to reflect differences in performance and uncertainty (Marmion et al., 2009). We suggest that a future avenue of research should focus on how the number of background points used to train component models influences the process by which we assign model weights during the consensus-building process like WA and mPCA.

5. Conclusion

Multi-scale ensemble habitat selection models have been steadily gaining popularity given their perceived superiority over individual modeling methods, but the effects of background points were previously unchecked. We assessed multiple combinations of models built with different numbers of background points to explore how the number of background points impacts the performance of ensemble models. We found that an ensemble of models, each trained with an optimal number of background points, outperformed ensembles of models trained with the same number of background points, although the differences in performance were small. Even though our four ensemble models performed similarly when assessed with AUC, for now we suggest that optimizing the number of background points for each component model of an ensemble model can help avoid any unnecessary negative impacts to model performance by a single component model built with suboptimal training data. However, RF outperformed all the ensemble models, suggesting that if researchers use a single modeling approach, they should consider using RF. It should be noted that we found RF to be more sensitive to the number of background points used to train the model than other modeling methods. However, researchers or practitioners might choose to use a single modeling approach that is found to be less sensitive to the number of background points, like a GLM or MaxEnt. In this situation, since the gain in performance is minimal, the decision on how many background points to use might be instead made to minimize computation time and answer time-sensitive questions related to conservation actions. Nevertheless, using an ensemble of models helps overcome any issues that any single model might have, as demonstrated by its relatively constant performance regardless of the number of background points used to train component models. As with many ecological problems, the decisions made during the modeling process demand careful consideration in the full context of the question at hand, since different questions often differ in their requirements from a model. In the future, we suggest evaluating more models across multiple species to investigate how background point selection might influence ensemble models in a variety of contexts.

Funding

This research was supported by a Joint Venture with the USDA Forest Service, Sponsor Award # 20-JV-11221633-152.

Declaration of Competing Interest

None of the authors have a conflict of interest to declare.

Data availability

The authors do not have permission to share data.

Acknowledgements

We thank the California Department of Fish and Wildlife for making northern spotted owl data available through the Spotted Owl Observations Database (SPOWDB). We thank the people from California Department of Fish and Wildlife, the United States Forest Service, Green Diamond Resource Company, and more who collected the data for their work. We also thank José-Juan Rodriguez for very helpful feedback on our figures.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2022.101914.

References

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. https://doi.org/10.1111/j.1365-2664.2006.01214.x.

Araújo, M.B., New, M., 2007. Ensemble forecasting of species distributions. Trends Ecol. Evol. 22 (1), 42–47. https://doi.org/10.1016/j.tree.2006.09.010.

Araújo, M.B., Whittaker, R.J., Ladle, R.J., Erhard, M., 2005. Reducing uncertainty in projections of extinction risk from climate change. Glob. Ecol. Biogeogr. 14 (6), 529–538. https://doi.org/10.1111/j.1466-822X.2005.00182.x.

Atzeni, L., Cushman, S.A., Bai, D., Wang, J., Chen, P., Shi, K., Riordan, P., 2020. Meta-replication, sampling bias, and multi-scale model selection: A case study on snow leopard (*Panthera uncia*) in western China. Ecol. Evol. 10 (14), 7686–7712. https://doi.org/10.1002/ece3.6492.

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. https://doi.org/10.1111/j.2041-210X.2011.00172.x.

Bates, J.M., Granger, C.W., 1969. The combination of forecasts. Oper. Res. Q. 20, 451–468. https://doi.org/10.2307/3008764.

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multi-Model Inference: A Practical Information-Theoretical Approach, Second edition. Springer, Berlin, Germany.

Chiaverini, L., Wan, H.Y., Hahn, B., Cilimburg, A., Wasserman, T.N., Cushman, S.A., 2021. Effects of non-representative sampling design on multi-scale habitat models: flammulated owls in the Rocky Mountains. Ecol. Model. 450, 109566 https://doi.org/10.1016/j.ecolmodel.2021.109566.

Coetzee, B.W.T., Robertson, M.P., Erasmus, B.F.N., Van Rensburg, B.J., Thuiller, W., 2009. Ensemble models predict important bird areas in southern Africa will become less effective for conserving endemic birds under climate change. Glob. Ecol. Biogeogr. 18 (6), 701–710. https://doi.org/10.1111/j.1466-8238.2009.00485.x.

Conservation International. 2022. Explore the Biodiversity Hotspots | CEPF. Retrieved June 29, 2022, from https://www.cepf.net/our-work/biodiversity-hotspots.

Cushman, S.A., Macdonald, E.A., Landguth, E.L., Malhi, Y., Macdonald, D.W., 2017. Multiple-scale prediction of forest loss risk across Borneo. Landsc. Ecol. 32 (8), 1581–1598. https://doi.org/10.1007/s10980-017-0520-0.

Dormann, C.F., McPherson, M., Araújo, J.B., Bivand, M., Bolliger, R., Carl, J., Davies, G. G., Hirzel, R., Jetz, A., Kissling, W. Daniel, Kühn, W., Ohlemüller, I., Peres-Neto, R. R., Reineking, P., Schröder, B., Schurr, F.B.M., Wilson, R., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30 (5), 609–628. https://doi.org/10.1111/j.2007.0906-7590.05171.x.

Elith, J., Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40 (1), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159.

Elith*, J., Graham*, H., Anderson, C.P., Dudík, R., Ferrier, M., Guisan, S., Hijmans, A.J., Huettmann, R., Leathwick, F.R., Lehmann, J., Li, A., Lohmann, J.G., Loiselle, L.A., Manion, B., Moritz, G., Nakamura, C., Nakazawa, M., McC, Y., Overton, M., Townsend, J., Peterson, A., Zimmermann, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29 (2), 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x.

Forsman, E.D., Anthony, R.G., Reid, J.A., Loschl, P.J., Sovern, S.G., Taylor, M., Biswell, B.L., Ellingson, A., Meslow, E.C., Miller, G.S., Swindle, K.A., Thrailkill, J.A., Wagner, F.F., Seaman, D.E., 2002. Natal and breeding dispersal of northern spotted owls. Wildl. Monogr. 149, 1–35. https://www.jstor.org/stable/3830803.

Gregory, A.W., Smith, G.W., Yetman, J., 2001. Testing for forecast consensus. J. Bus. Econ. Stat. 19 (1), 34–43.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135 (2), 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9.

Hao, T., Elith, J., Lahoz-Monfort, J.J., Guillera-Arroita, G., 2020. Testing whether ensemble modelling is advantageous for maximising predictive performance of species distribution models. Ecography 43 (4), 549–558. https://doi.org/10.1111/ecog.04890.

Hegel, T.M., Cushman, S.A., Evans, J., Huettmann, F., 2010. Current state of the art for statistical modelling of species distributions. In: Cushman, S.A., Huettmann, F. (Eds.), Spatial Complexity, Informatics, and Wildlife Conservation. Springer, Japan, pp. 273–311. https://doi.org/10.1007/978-4-431-87771-4_16.

Kumar, S.U., Maini, P.K., Chiaverini, L., Hearn, A.J., Macdonald, D.W., Kaszta, Ż., Cushman, S.A., 2021. Smoothing and the environmental manifold. Ecol. Inform. 66, 101472 https://doi.org/10.1016/j.ecoinf.2021.101472.

LANDFIRE, 2017. Existing Vegetation Type Layer, LANDFIRE 2.0.0, U.S. Department of the Interior, Geological Survey, and U.S. Department of Agriculture. Accessed 28 October 2021 at. http://www.landfire/viewer.

Liu, C., White, M., Newell, G., 2013. Selecting thresholds for the prediction of species occurrence with presence-only data. J. Biogeogr. 40 (4), 778–789. https://doi.org/10.1111/jbi.12058.

Liu, C., Newell, G., White, M., 2016. On the selection of thresholds for predicting species occurrence with presence-only data. Ecol. Evol. 6 (1), 337–348. https://doi.org/10.1002/ece3.1878.

Liu, C., Newell, G., White, M., 2019. The effect of sample size on the accuracy of species distribution models: considering both presences and pseudo-absences or background sites. Ecography 42 (3), 535–548. https://doi.org/10.1111/ecog.03188.

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R.K., Thuiller, W., 2009. Evaluation of consensus methods in predictive species distribution modelling. Divers. Distrib. 15 (1), 59–69. https://doi.org/10.1111/j.1472-4642.2008.00491.x.

McGarigal, Kevin, Wan, Ho Yi, Zeller, Kathy, Timm, Brad, Cushman, Samuel, et al., 2016. Multi-scale habitat selection modeling: a review and outlook. Landscape Ecol. 31, 1161–1175. https://doi.org/10.1007/s10980-016-0374-x.

Mohammadi, A., Almasieh, K., Nayeri, D., Adibi, M.A., Wan, H.Y., 2022. Comparison of habitat suitability and connectivity modelling for three carnivores of conservation concern in an Iranian montane landscape. Landsc. Ecol. 37 (2), 411–430. https://doi.org/10.1007/s10980-021-01386-5.

Ohmann, J.L., Gregory, M.J., 2002. Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, USA. Can. J. For. Res. 32 (4), 725–741. https://doi.org/10.1139/x02-011.

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190 (3–4), 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026.

Ramirez-Reyes, C., Nazeri, M., Street, G., Jones-Farrand, D.T., Vilella, F.J., Evans, K.O., 2021. Embracing ensemble species distribution models to inform at-risk species status assessments. J. Fish and Wildlife Manag. 12 (1), 98–111. https://doi.org/10.3996/JFWM-20-072.

Rew, J., Cho, Y., Moon, J., Hwang, E., 2020. Habitat suitability estimation using a two-stage ensemble approach. Remote Sens. 12 (9), 1475. https://doi.org/10.3390/rs12091475.

Rupp, D.E., Daly, C., Doggett, M.K., Smith, J.I., Steinberg, B., 2022. Mapping an observation-based global solar irradiance climatology across the conterminous United States. J. Appl. Meteorol. Climatol. 61 (7), 857–876. https://doi.org/10.1175/JAMC-D-21-0236.1.

Shabani, F., Kumar, L., Ahmadi, M., 2016. A comparison of absolute performance of different correlative and mechanistic species distribution models in an independent area. Ecol. Evol. 6 (16), 5973–5986. https://doi.org/10.1002/ece3.2332.

Sofaer, H.R., Hoeting, J.A., Jarnevich, C.S., 2019. The area under the precision-recall curve as a performance metric for rare binary events. Methods Ecol. Evol. 10 (4), 565–577. https://doi.org/10.1111/2041-210X.13140.

Thuiller, W., 2004. Patterns and uncertainties of species' range shifts under climate change. Glob. Chang. Biol. 10 (12), 2020–2027.

Thuiller, W., Georges, D., Gueguen, M., Engler, R., Breiner, F., 2021. biomod2: Ensemble Platform. For Species Distribution Modeling (3.5.1) [Computer Software]. https://CRAN.R-project.org/package=biomod2.

Torres, J., Brito, J.C., Vasconcelos, M.J., Catarino, L., Gonçalves, J., Honrado, J., 2010. Ensemble models of habitat suitability relate chimpanzee (*pan troglodytes*) conservation to forest and landscape dynamics in Western Africa. Biol. Conserv. 143 (2), 416–425. https://doi.org/10.1016/j.biocon.2009.11.007.

United States Fish, Wildlife Service, [USFWS]., 2011. Revised recovery plan for the Northern spotted owl (Strix occidentalis caurina). U.S. Fish and Wildlife Service, Portland, Oregon.

United States Forest Service [USFS]. (2017). Ecological subregions: sections and subsections for the conterminous United States.

Valavi, R., Elith, J., Lahoz-Monfort, J.J., Guillera-Arroita, G., 2019. blockCV: an r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods Ecol. Evol. 10 (2), 225–232. https://doi.org/10.1111/2041-210X.13107.

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J., Elith, J., 2021. Predictive performance of presence- only species distribution models: a benchmark study with reproducible code. Ecol. Monogr. 92 (1), e01486 https://doi.org/10.1002/ecm.1486.

VanDerWal, J., Shoo, L.P., Graham, C., Williams, S.E., 2009. Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know? Ecol. Model. 220 (4), 589–594. https://doi.org/10.1016/j.ecolmodel.2008.11.010.

Weisel, L.E., 2015. Northern Spotted Owl and Barred Owl Home Range Size and Habitat Selection in Coastal Northwestern California. Master's thesis. Humboldt State University, Scholarworks.

Wiens, John, 1989. Spatial Scaling in Ecology. Func. Ecol. 3 (4), 385–397. https://doi.org/10.2307/2389612.

Zeller, K.A., Cushman, S.A., Van Lanen, N.J., Boone, J.D., Ammon, E., 2021. Targeting conifer removal to create an even playing field for birds in the Great Basin. Biol. Conserv. 257, 109130 https://doi.org/10.1016/j.biocon.2021.109130.

Zhu, G., Fan, J., Peterson, A.T., 2021. Cautions in weighting individual ecological niche models in ensemble forecasting. Ecol. Model. 448, 109502 https://doi.org/10.1016/j.ecolmodel.2021.109502.