# The effect of sample size on the accuracy of species distribution models: considering both presences and pseudo-absences or background sites

Canran Liu, Graeme Newell and Matt White

Arthur Rylah Institute for Environmental Research, Department of Environment, Land, Water and Planning, 123 Brown Street, Heidelberg, Victoria 3084, Australia

**Corresponding author:** Canran Liu, Arthur Rylah Institute for Environmental Research, Department of Environment, Land, Water and Planning, 123 Brown Street, Heidelberg, Victoria 3084, Australia.
E-mail: canran.liu@delwp.vic.gov.au

**Decision date:** 05-Jul-2018

---

**Abstract**

Most high-performing species distribution modelling techniques require both presences, and either absences or pseudo-absences or background points. In this paper, we explore the effect of sample size, towards developing improved strategies for modelling. We generated 1,800 virtual species with three levels of prevalence using ten modelling techniques, while varying the number of training presences (NTP) and the number of random points (NRP representing pseudo-absences or background sites). For five of the ten modelling techniques we built two versions of models: one with an equal total weight (ETW) setting where the total weight for pseudo-absence is equivalent to the total weight for presence, and another with an unequal total weight (UTW) setting where the total weight for pseudo-absence is not required to be equal to the total weight for presence. We compared two strategies for NRP: a small multiplier strategy (i.e. setting NRP at a few times as large as NTP), and a large number strategy (i.e. using numerous random points). We produced ensemble models (by averaging the predictions from 30 models built with the same set of training presences and different sets of random points in equivalent numbers) for three NTP magnitudes and two NRP strategies. We found that model accuracy altered as NRP increased with four distinct patterns of performance: increasing, decreasing, arch-shaped and horizontal. In most cases ETW improved model performance. Ensemble models had higher accuracy than the corresponding single models, and this improvement was pronounced when NTP was low. We conclude that a large NRP is not always an appropriate strategy. The best choice for NRP will depend on the modelling techniques used, species prevalence and NTP. We recommend building ensemble models instead of single models, using the small multiplier strategy for NRP with ETW, especially when only a small number of species presence records are available.

## Introduction

Species distribution models (SDMs) have been widely used to assess the impact of climatic change on species distributions and the impact of species invasion, to select sites for species reintroduction, and to establish conservation priorities. Various techniques have been developed to construct SDMs (Elith et al. 2006, Warton and Shepherd 2010, Chakraborty et al. 2011, Aarts et al. 2012). Some methods use only presence data, which are called profile methods (Robertson et al. 2001), e.g. BIOCLIM, DOMAIN and LIVES (Li and Hilbert 2008), while others use both presence and absence data, which are called group discrimination methods (Robertson et al. 2001), e.g. generalised linear model (GLM) and generalised additive model (GAM). Other methods use presence data and background data, including ecological niche factor analysis (ENFA, Hirzel et al. 2002) and Maxent (Phillips et al. 2008). Since these latter methods do not require absence data, they have also been considered as profile techniques (Wisz and Guisan 2009, Bean et al. 2011). In general, the group discrimination methods have higher performance than profile methods apart from Maxent (Elith et al. 2006). Variability in the performance of modelling techniques, as well as the influence of the species presence data (e.g. Grenouillet et al. 2011), have led a number of researchers to recommend an ensemble modelling approach (Thuiller 2004, Araújo and New 2007, Marmion et al. 2009, Thuiller et al. 2009, Grenouillet et al. 2011, Comte and Grenouillet 2013).

The occurrence data for many species are sourced from museums and other databases, which generally collate species' presence data, but not absence data. The use of group discrimination methods requires the generation of some form of pseudo-absence data as a surrogate for true absence data. Pseudo-absences may be generated using several approaches (Graham et al. 2004, Lobo and Tognelli 2011), e.g. selecting random points from the study area (Stockwell and Peters 1999, Ferrier et al. 2002), using a profile model to filter the points within the study area (Engler et al. 2004, Lobo and Tognelli 2011, Liu et al. 2013a, 2013b), selecting pseudo-absences that have been biased in a way as species records (Dudík et al. 2005), target group approach (Phillips et al. 2009), and three-step technique (Senay et al. 2013). Using random points within the geographic region of interest is the simplest and most widely used of all these approaches.

Some clarification is required around the terms: absences, pseudo-absences and background sites. Absences (or true absences for emphasis) are the sites where environmental conditions are not suitable for the species or are suitable sites to which the species could not be dispersed. Pseudo-absences are

sites chosen to represent absences, although we don't know if they are true absences or not. Background sites are chosen without regard to the distribution of presence/absence locations, which can be all the sites within the study area or a random sample of the sites. Background sites should be used with Maxent and ENFA, and pseudo-absences, which include background sites as a special type, should be used with the group discrimination methods. Since random points are both background sites and a type of pseudo-absences, they can be used with all the methods.

Several factors have been found to affect the performance of SDMs, including sample size (Pearce and Ferrier 2000, Stockwell and Peterson 2002, Kadmon et al. 2003, McPherson et al. 2004, Pearson et al. 2007, Jiménez-Valverde et al. 2009, Santika and Hutchinson 2009, Chefaoui et al. 2011, Dupin et al. 2011, Barbet-Massin et al. 2012, Hanberry et al. 2012, Loe et al. 2012), species characteristics (including range size, spatial extent, clumping, dispersal ability, marginality, tolerance, and prevalence − the ratio of the area a species occupies to the size of the study area, e.g. McPherson et al. 2004, Segurado and Araújo 2004, Reese et al. 2005, Hernandez et al. 2006, Franklin et al. 2009, Marmion et al. 2009, Chefaoui et al. 2011), and scale (Thuiller et al. 2003, Seo et al. 2009, Guisan et al. 2007, Gottschalk et al. 2011, Pittman and Brown 2011). The potential effects of sample size have attracted considerable attention, particularly relating to circumstances where there are a limited number of occurrence records available for modelling certain species with significant conservation concerns (Wisz et al. 2008). In these circumstances researchers require protocols to have confidence in selecting an appropriate sample size and strategy to develop models with an acceptable performance.

Some studies have focused on the effect of presence sample size (including the above-mentioned studies), and generally found that model accuracy increased as the sample size of presence data increased. This result is rational for profile models, since only presence information is included in the models. However, when two components (both presences and absences or pseudo-absences) are involved in the modelling for group discrimination models, only considering the sample size of presence data is an inadequate strategy. The sample size for both presence and absence or pseudo-absence elements needs to be considered together.

While the paucity of presence data for many species restrict the presence sample, we can generate as many pseudo-absences or background sites as we want for many species, particularly when allocating random points (i.e. up to the total number of cells within the study area, see Chefaoui et al. 2016 for an exception). This approach however does not clarify the question of deriving an optimal

number of pseudo-absences or background sites required for the development of a high-quality model. In GARP (Stockwell and Peters 1999) and Maxent (Phillips et al. 2006) 1250 and 10000 random points have respectively been used as default settings. These settings have subsequently been followed by many users of these software (e.g. Gibson et al. 2007). Large numbers of pseudo-absences have also been used with other modelling techniques, e.g. 1000 pseudo-absences for GAM (Ferrier et al. 2002), 1000 and 3000 random points for boosted regression tree (BRT) and GLM (Elith and Graham 2009), 5000 points (with some restriction and not completely random) for GLM (Gibson et al. 2007), and 10000 random points for BRT, GLM, GAM, generalised dissimilarity model (GDM), multivariate adaptive regression splines (MARS) (Elith et al. 2006, Wisz et al. 2008) and random forest (RF) (Hernandez et al. 2008).

Not all researchers have used large numbers of pseudo-absences. Some have adopted specific ratios of the number of pseudo-absences or background sites to the number of presences. Equal numbers of pseudo-absences or background sites as presences have been used for GLM and GARP models by Stockwell and Peterson (2002), for eight modelling techniques (including artificial neural network – ANN, BRT, classification tree analysis – CTA, GLM, GAM, MARS, mixture discriminant analysis – MDA and RF) by Capinha and Anastácio (2011), and for BRT, GLM, MARS, GARP and Maxent by Mateo *et al*. (2012). A ratio of 10 times more pseudo-absences than presences has been used for GLM by Chefaoui and Lobo (2008) and for BRT, GAM, GLM and RF by Qiao et al. (2015), and a two times ratio of pseudo-absences to presences has been used for RF and GAM by Liu et al. (2013a, 2013b).

Few studies to date have investigated the effect of the number of pseudo-absences and background sites on model accuracy. Phillips et al. (2008) built Maxent models with the data for 226 species (with the number of presences ranging from 2 to 5822) from six regions of the world by using random points scaled in a geometric progression from 63 to 256000. They demonstrated that model accuracy (measured by the area under the receiver's operating characteristic curve – AUC) increased substantially as number of random points increased, reaching a plateau after 8000 random points. Lobo and Tognelli (2011) used a virtual species with prevalence of 3.5%, and built GAMs with 10 presences and pseudo-absences at three frequencies (10, 100 and 1000). Results from this study also identified that model accuracy (AUC) increased as the number of pseudo-absences increased, and suggested that where limited presence data are available, 100:1 of pseudo-absences to presences be employed.

Barbet-Massin et al. (2012) studied the effect of the number of pseudo-absences and weighting schemes by using seven modelling techniques, two virtual species with similar prevalence, three levels of the number of presences (30, 100, 300) and three levels of the number of pseudo-absences (100, 1000, 10000). They found both the number of pseudo-absences and weighting schemes had significant effects on model accuracy.

It is worth noting from the above studies that Lobo and Tognelli (2011) used one virtual species and one modelling technique (GAM) while Phillips et al. (2008) used 226 species and a single modelling technique (Maxent), without investigating the effect of species prevalence, and although Barbet-Massin et al. (2012) used seven modelling techniques and two weighting schemes, they only used two virtual species with similar prevalence. Ideally many or all the relevant factors and interactions (i.e. modelling techniques, weighting schemes, species prevalence, and sample size for both presences and pseudo-absences or background sites) could be considered in a structured manner to form a more complete view of their individual and collective effects upon model accuracy. In this paper, we have selected ten modelling techniques (and five with two weighting schemes) to build both single models and (within-method) ensemble models (with consensus method), and have used virtual species with varying levels of prevalence and disparate numbers of training presences to investigate the response of model accuracy to both the number of presences and the number of pseudo-absences or background sites.

## Methods

### Modelling techniques

Ten modelling techniques were investigated in this study, including ANN, BRT, ENFA, GAM, GLM, MARS, Maxent, MDA, RF and support vector machine (SVM). ENFA and Maxent were included as both of these profile techniques (Wisz and Guisan 2009, Bean et al. 2011) also use background data.

The ten techniques examined in this study have been widely used in species distribution modelling. GLM, GAM, ANN and ENFA have been used in this pursuit for more than a decade (Guisan and Zimmermann 2000). Maxent has become a popular technique due to its excellent performance (e.g. Elith et al. 2006) and accessibility. Elith et al. (2006) also demonstrated that BRT was also among the top performing techniques, and that the performance of MARS was comparable to

GLM and GAM. Similarly, RF, SVM and MDA have been adopted as they also perform well on typical SDM problems [e.g. Cutler et al. (2007) and Liu *et al*. (2013a, 2013b) for RF, and Guo et al. (2005) and Drake et al. (2006) for SVM, Marmion et al. (2009) and Capinha and Anastácio (2011) for MDA]. The specialized SDM software Maxent (Phillips et al. 2006) and the R packages BIOMOD (Thuiller et al. 2009), dismo (Hijmans and Elith 2013) and sdm (Naimi and Araujo 2016) have facilitated the implementation and use of these techniques.

All analyses in this study were conducted within the R environment. ENFA was carried out with our own code according to Hirzel et al.'s (2002) algorithm. The other techniques were implemented with R packages, including nnet for ANN, gbm for BRT, mgcv for GAM, stats for GLM, earth for MARS, dismo interface with Phillips' Java program for Maxent, mda for MDA, randomForest for RF, and kernlab for SVM. All the models were developed using the default setting of the software packages, which is a common practise for many researchers (e.g. Phillips et al. 2009, Araújo et al. 2011, Qiao et al. 2015). Alternate versions of models were also built using five techniques (ANN, BRT, GAM, GLM and SVM), where the total weight between presences and pseudo-absences was equal. For this investigation we termed this an equal total weight (ETW) setting, and the default settings of the packages to be termed unequal total weight (UTW) setting.

**Generation of virtual species**

For this study we used virtual species distributed in a 250 km × 250 km region across central western Victoria, Australia. This study area spans a range of distinct geomorphological and climatic contexts from near sea level to 1100 m in altitude. Eighteen environmental variables, including bioclimatic, topographic and radiometric variables, were used to extract six principal components ( $x_j$, $j = 1,2,\cdots,6$ ) using principal component analysis (PCA), which accounted for more than 85% of the total variation. Environmental data had an original primary resolution of 75 m × 75 m, but were rescaled to 1 km × 1 km, creating a total of $n_T = 62500$ cells. See Liu et al. (2013b) for more details of the variables.

Two approaches were used to create virtual species: probabilistic and threshold (Meynard and Kaplan 2013). Both approaches have been employed previously in threshold selection studies (Liu et al. 2013a, Liu et al. 2016). From our experience, it is very difficult to create virtual species with a specified prevalence, when using the probabilistic approach. As this is an important factor affecting

model performance, the threshold approach was adopted in this study for simplicity. Virtual species were simulated as follows. For each species, random values were selected: $a_j$ ( $j = 0,1,\cdots,6$ ) and $b_{jk}$ $(j,k = 1,2,\cdots,6; j \le k)$, which are uniformly distributed on the interval (-1,1). For site $i$ ( $i = 1,2,\cdots,n_T$ ) with environmental data $X_i = (x_{1i}, x_{2i}, \cdots, x_{6i})$, the suitability for the species was calculated as

$$P(X_i) = \frac{1}{1+e^{-f(X_i)}},$$

where

$$f(X_i) = a_0 + \sum_{j=1}^{6} a_j x_{ji} + \sum_{j \le k}^{6} b_{jk} x_{ji} x_{ki}$$

To make the prevalence of the species a specific value $p$ , the $n_T p$ cells with the highest suitabilities were labelled as presences, and all the other cells absences. Examples of the simulated species are shown in Supplementary material Appendix 1 Fig. A1.

**Dataset creation and model building and testing**

To investigate the effect of the number of pseudo-absences or background points on model accuracy we simulated the distribution of 200 species for each of the three levels of species prevalence ($p$): 0.05 (rare), 0.25 (common) and 0.75 (very common), and set three levels for the number of training presences (NTP): 20, 160 and 640. For each level of NTP we set ten levels for the number of random points (NRP): 10, 20, 40, 80, 160, 320, 640, 1280, 2560 and 5120. For each level of prevalence, after a species was simulated, a test dataset with 3000 data points was assembled, which was composed of randomly sampled $3000p$ presences and $3000(1-p)$ absences, where $p$ was species prevalence. For each level of NTP, we randomly selected NTP presences from those after removing the test presences from all the presences. For each level of NRP, we randomly sampled NRP points from all the points ($n_T$) in the whole study area. With these presences and random points we built a model using one of the ten techniques, applied the model to the test dataset, and made predictions for the test data points. While some researchers remove random points where they coincided with training presence data points, in this study we retained the complete set of random points since pseudo-absence selection was not the primary aim of this study.

To investigate the effect of NTP on model accuracy we simulated the distribution of an additional 200 species for each of the above three levels of species prevalence. Nine levels for NTP were set: 10,

20, 40, 80, 160, 320, 640, 1280 and 2560. Two strategies were used for NRP: small multiplier and large number. For the former we used twice as many NRP as NTP, following Liu et al. (2013a, 2013b). For the latter we set NRP = 5120 (instead of 10000 as is often used, to facilitate the calculation). This "large number" strategy is commonly used, especially for Maxent models. The details of the simulation were the same as described earlier.

We also investigated the effect of sample size on the accuracy of ensemble models by simulating an additional 200 species for each of the three levels of prevalence. We set three levels for NTP (20, 160 and 640) and used the same two strategies for NRP as detailed above. After a species was created, a test dataset was compiled using the approach described earlier. For each level of NTP we compiled a set of presences as training presences, randomly selected a specified number of sites using either of the two strategies for NRP, built a model with each of the ten modelling methods, and then applied the model to the test dataset. This process was repeated 30 times (but each time the training presences were retained and only the random points were altered), and the 30 sets of predictions from the same modelling method were then averaged into an ensemble model. With this approach the term ensemble as used in this study is as a within-method ensemble. One (randomly chosen) of the 30 models was kept as the single model. This allowed for a comparison between "single" (i.e. conventional) models and those developed as 'ensemble' models.

Five accuracy measures were calculated to assess model performance. These included AUC, point biserial correlation coefficient (Rpb), true skill statistic (TSS), sensitivity (Se) and specificity (Sp) (Liu et al. 2011). To calculate the last three metrics, a threshold obtained by maximizing the sum of Se and Sp (Liu et al. 2005), was applied to the model predictions to transform them from continuous to binary forms. As statistical tests are not encouraged for simulation modelling results (White et al. 2014), we used the 2.5th, 50th (i.e. median) and 97.5th percentiles of the predictions to compare different scenarios. If the 2.5th percentile of the predictions for scenario A is above the 97.5th percentile of the predictions for scenario B, scenario A is considered to have better performance than scenario B. This is a stringent test. But this occurred infrequently. Therefore, we also define a less stringent test. Where the 2.5th percentile of the predictions for scenario A is not above the 97.5th percentile of the predictions for scenario B, if at least one of the above three percentiles of the predictions for scenario A is above its counterpart for scenario B and the other percentiles for scenario A are not below their counterparts for scenario B, then scenario A is considered to have better performance than scenario B.

Despite the reservations of some, formal statistical analyses are still helpful in summarizing the results of simulations. To characterize the effect of NRP on the accuracy of models, we conducted quantile regression analysis by fitting three models: (1) $AUC = b\ln(NRP) + c$, (2) $AUC = a[\ln(NRP)]^2 + b\ln(NRP) + c$, and (3) $AUC = a(NRP)^b + c = ae^{bln(NRP)} + c$ for each combination of the factors investigated. We used two functions rq and nlrq in the R package quantreg for these analyses with the parameter $tau = 0.5$, i.e. we focused on estimating the conditional median. We used AIC to select the best models. When we identified the pattern of the response of the accuracy to NRP, we also considered the absolute effect size. At least one percent difference was considered a real change in accuracy. For example, if an increasing or decreasing pattern was to be identified, the difference in accuracy between the two ends of the curve must be no less than one percent; if an arch-shaped pattern was to be identified, the difference in accuracy between the maximum value and both ends of the curve must be no less than one percent.

**Data deposition**

Data available from the Mendeley Data: < http://dx.doi.org/10.17632/kyy8f8w79j.1> (Liu et al. 2018).

# Results

The effects of NRP on the accuracy of models were found to vary considerably when using different modelling techniques, levels of species prevalence, levels of NTP, and weighting schemes (Figs. 1-2, Supplementary material Appendix 2 Figs. A1-A21). These results could be summarized into four types or patterns of effects on model accuracy as a function of NRP (see Fig.3 and Supplementary material Appendix 2 Fig. A22 for some examples). These were termed increasing, decreasing, arch-shaped, and horizontal (i.e. no effect). It was common that more than one effect response was observed for a single modelling technique.

Model accuracy increased with NRP in many situations when measured with AUC (Supplementary material Appendix 2 Tables A1-A2), especially for ENFA, Maxent, GLM and BRT, MDA for common and very common species, MARS and ANN for larger NTP, and GAM with ETW. In most of these situations accuracy reached an asymptote when NRP became a few times as large as NTP, except for GAM and GLM with smaller NTP where NRP was required up to 16 times as large as NTP. For RF and SVM the response curves were often arch-shaped. The response curves for ANN with a

smaller NTP displayed either a decreasing or horizontal pattern. In these situations, the accuracy peaked when NRP became a few times as large as NTP.

The response patterns were generally consistently displayed when being assessed with TSS, Se and Sp in comparison to AUC. Accuracy increased as NRP increased for ENFA, Maxent, and ANN, BRT, GAM, GLM and SVM with ETW when using Rpb as the accuracy measure. For most other situations the response curves were arch-shaped.

The strategies for allocating NRP (i.e. the small multiplier strategy versus the large number strategy) were examined using nine levels of training presences (Fig. 4, Supplementary material Appendix 3 Figs. A23-A40). The small multiplier strategy for NRP produced models with higher accuracy for RF, SVM (with UTW), ANN (except very common species), BRT (with UTW except very common species), and MARS (rare species). The large number strategy produced models with a higher accuracy only for ENFA, MDA and Maxent (for very common species), and GAM and GLM (except UTW with smaller NTP).

Model accuracy (AUC) increased as NTP increased for all the combinations of modelling technique, prevalence and weight setting. The accuracy of the models reached their respective asymptotes before NTP increased to the highest level for all the modelling techniques (except RF, ANN and SVM with large number strategy and very common species). However, the NTP required for the models to reach these asymptotes varied between modelling techniques, with alternate NRP strategies, and with levels of prevalence. ENFA models reached the asymptotes of accuracy much earlier than other methods. In general, the more elevated a species' prevalence, the larger the NTP that was required for the models to approach asymptotes of accuracy.

Ensemble models were better than their individual counterparts in most cases (Figs. 5-6, Supplementary material Appendix 4 Figs. S41-S48). For ANN this was true for all situations. The difference was obvious for GAM, GLM, MARS and MDA for rare species and smaller NTP. For example, GLM for NTP = 20 and both weight settings yielded the only two cases that met the requirement for the stringent test, i.e. the 2.5th percentiles of the predictions for ensemble models were above the 97.5th percentiles of the predictions for the corresponding single models, while the other cases only met the requirement for less stringent test. Ensemble models were also better than single models for Maxent (with small multiplier strategy for NRP) and RF.

In many comparisons ensemble models with the two weight settings had similar accuracy. In some situations, those with ETW had a higher accuracy than counterparts with UTW. This was particularly pronounced for SVM. In almost all situations ensemble models with small multiplier strategy for NRP had higher accuracy than counterparts with large number strategy for NRP. The exception to this was for ENFA, where almost no difference was observed between single and ensemble models using a small multiplier or large number strategy for NRP.

## Discussion

### Considerations for total weight setting

The equal total weight (ETW) setting has been used with a variety of techniques within a range of studies, e.g. Ferrier et al. (2002) using GAM, Elith et al. (2006) using BRT, BRUTO, GAM and MARS, Elith and Graham (2009) using BRT and GLM, and Qiao et al. (2015) using BRT, GAM and GLM. Additionally, Elith and Graham (2009) compared the performance of models with ETW against unequal total weight (UTW) setting, and found that the effect of weighting varied for different techniques. We also have observed a similar effect in this study.

Although ETW generally displayed higher performance than UTW this is not always the case. ANN is such an exception when NTP was small and prevalence was low to moderate. In these situations, the accuracy-NRP (number of random points) curve was still decreasing while it was increasing for other methods. Our study has also clearly demonstrated some benefits of adopting an ETW approach, as it provides smoother accuracy-NRP curves, particularly for the ANN, GLM and SVM models. This means that with ETW models became more stable. Second, with ETW the accuracy-NRP curves for GLM and SVM models changed from decreasing or arch-shaped to nearly increasing or horizontal. This makes it easier to select an optimal NRP, since it supports the large number strategy, so a wide range of NRP can be used. On the contrary, with UTW sometimes the curves are sharply peaked and the peaks are dependent on both prevalence and NTP, e.g. for SVM models with very common species. In this case, if the optimal NRP is incorrectly identified, the accuracy of the resulted models will decrease abruptly. For this reason, the use of ETW setting is generally encouraged.

### Considerations for the number of random points

By systematically altering the number of random points (NRP), we have identified detailed patterns in the responses of the accuracy models in response to the NRP. We have observed at least four types of patterns for the NRP effect, including increasing, decreasing, arch-shaped and horizontal responses. Apart from the increasing pattern, all other patterns do not encourage the use of a very large NRP. Even for the techniques displaying an increasing response, the use of a large NRP would not provide further advantages beyond the point where NRP becomes a few times greater than NTP. The exception to this is for GAM and GLM at small NTP, where NRP needs to be up to 16 times greater than NTP for the accuracy to approach an asymptote. Some of these results are consistent with Barbet-Massin et al. (2012). For example, they have recommended a large NRP (i.e. our large number strategy) for GAM and GLM when ETW was adopted, and an equal number of pseudo-absences and available presences (similar to our small multiplier strategy) for RF. However, there are some inconsistencies between our results and theirs. For example, they have suggested the small multiplier strategy for MARS and MDA, but in this study we found that this is only appropriate for rare species. This inconsistency may be due to the prevalence of the virtual species, the level of presences, spatial extent and spatial resolution used in the two studies. The prevalence of their two virtual species is similar and is less than 25% as judged from their maps. If this is taken into account, there is no inconsistency between the two studies. Therefore, in order to reach a reliable conclusion, a wide range of the parameter values should be considered in the simulation.

For the Maxent method we identified similar model accuracy (AUC) response patterns to NRP as with Phillips et al. (2008) study, i.e. accuracy roughly increases with NRP. However, we also found that the large number strategy is only better for very common species with NTP < 160, and outside these settings, the advantage of the large number strategy didn't exist (Fig. 3). This couldn't be revealed in Phillips et al. (2008) due to the design that they didn't differentiate varying levels of NTP and prevalence.

**Consideration of number of training presences**

This study has reconfirmed that model accuracy increases with the number of training presences (e.g. Stockwell and Peterson 2002; Wisz et al. 2008). This is a general pattern for all the modelling techniques studied, provided sufficient pseudo-absences are used in the modelling. Otherwise, other patterns may emerge, e.g. for ANN and MARS for very common species with NRP < 20 and NTP = 20 (e.g. prevalent but poorly surveyed, Figs. 1-2). This means that the above general pattern is not a

universal one. It is a true pattern for profile models requiring only presence data. But for group discrimination models, when the training data are derived from two separate samples (in the presence-only situation: one for presences and the other for pseudo-absences), model accuracy is affected by several factors related to the training dataset itself (here we do not consider modelling techniques): (1) sample size (for presences and for pseudo-absences), (2) the ratio of the number of the presences to the number of the pseudo-absences, and (3) the true prevalence of a species, which is hidden in the random points. When pseudo-absences are used to train models, the accuracy of the models may decrease as the NTP increases even when the NRP remains fixed but at low level. For routine modelling settings the general pattern is maintained.

The above general pattern encourages the maximal use of observations for building models, which is consistent with the findings of other studies (Stockwell and Peterson 2002, Hernandez et al. 2006, Wisz et al. 2008, Hanberry et al. 2012). We have also observed that the accuracy of many models reached their asymptotes above some specific NTP. But this does not prevent researchers from sourcing and supporting the collection of more species field observations. The reason is that to reach the asymptotes of accuracy hundreds of representative presences are needed. Even though many species have hundreds even thousands of records in the existing databases, these collective sets of observations are almost suffering from survey bias, and are therefore typically far from representative. Therefore, keeping collection of records and using them in modelling is still encouraged.

**Single models versus ensemble models**

Ensemble models, especially with consensus methods, typically refer to approaches that employ multiple modelling techniques (Araújo and New 2007, Marmion et al. 2009, Thuiller et al. 2009, Grenouillet et al. 2011). In this study, they have been used within modelling techniques. We have shown that even within modelling techniques, ensemble models still outperform their corresponding single models. This is consistent with previous studies using cross-technique ensemble modelling (Araújo et al. 2005, Marmion et al. 2009). The exception to this is for ENFA models, where the ensemble models and single models have similar accuracies. However, the effectiveness of ensemble modelling differs among the modelling techniques and data characteristics. For example, ensembling is more effective for ANN than for other techniques, and it is also more effective when fewer training presences are available and the small multiplier NRP strategy is used than in other contexts. It is interesting that either small multiplier NRP strategy or large number NRP strategy is better for single

models, small multiplier strategy is almost always better for ensemble models. This has an additional advantage in practise. That is, by adopting this ensemble modelling approach the computing burden will not increase too much and it may even decrease. The reason is that we do not need many models within an ensemble. From our experiments, we estimate that in most contexts, 30 models for an ensemble are adequate for the resultant model to approach an accuracy asymptote, and Barbet-Massin et al. (2012) even suggested using 10 models in an ensemble.

## Conclusions

In this study, we found that model accuracy altered in four patterns as the NRP increased: increasing, decreasing, arch-shaped and horizontal, and using ETW made the response curve change from decreasing and arch-shaped to nearly increasing or horizontal for some techniques (e.g. SVM and GLM). But for ANN models with ETW and small NTP, the response curves remained decreasing. The increasing response pattern tends to favour the use of a large number of random points, but the other three patterns do not favour this, and the small multiplier strategy (i.e. taking NRP a few times as large as NTP) is recommended in such instances. Even for the increasing pattern, accuracy approaches an asymptote when NRP reaches a few times as large as NTP (with the exception of GAM and GLM for small NTP which need 16 times), and deploying a larger NRP does not enhance performance. Ensemble models almost always perform better than their corresponding single models for a given NRP strategy and weighting scheme, especially when fewer presences are available. Ensemble models with the small multiplier NRP strategy almost always produce better models than with the large number NRP strategy for both weighting schemes. Weighting schemes do influence model performance, and ETW is better than UTW in most cases. Hence, ensemble modelling with ETW and small NRP strategy is generally encouraged, especially when fewer presences are available. These findings provide a set of heuristics for choosing appropriate numbers of random points and weight settings under different situations (rare versus common species and small versus large NTP) for both single models and ensemble models separately (Table 1). We hope these findings will assist modellers in making more informed decisions in the development of their models.

*Acknowledgements* – We thank Alan Robley, the subject editor and anonymous reviewers for their thoughtful comments. The authors are supported by the funding from Biodiversity Division, Department of Environment, Land, Water and Planning, Victoria, Australia.

## References

Aarts, G. et al. 2012. Comparative interpretation of count, presence–absence and point methods for species distribution models. – Methods in Ecology and Evolution 3: 177−187.

Araújo, M. B. and New, M. 2007. Ensemble forecasting of species distributions. – Trends Ecol. Evol. 22: 42–47.

Araújo, M.B. et al. 2011. Climate change threatens European conservation areas. – Ecology letters 14: 484−492.

Bean, W.T. et al. 2012. The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. – Ecography 35: 250−258.

Barbet-Massin, M. et al 2012. Selecting pseudo-absences for species distribution models: how, where and how many? – Methods in Ecology and Evolution 3: 327-338.

Capinha, C. and Anastácio, P. 2011. Assessing the environmental requirements of invaders using ensembles of distribution models. – *Diversity and Distributions* **17**: 13–24.

Chakraborty, A. et al. 2011. Point pattern modelling for degraded presence-only data over large regions. – Journal of the Royal Statistical Society. Series C, Applied Statistics 60: 757−776.

Chefaoui, R.M. and Lobo, J.M. 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. – Ecological Modelling 210: 478–486.

Chefaoui, R.M. et al. 2011. Effects of species' traits and data characteristics on distribution models of threatened invertebrates. – Animal Biodiversity and Conservation 34: 229–247.

Chefaoui, R. M. et al. 2016. Large-scale prediction of seagrass distribution integrating landscape metrics and environmental factors: the case of Cymodocea nodosa (Mediterranean–Atlantic). – Estuaries and Coasts 39: 123–137.

Comte, L. and Grenouillet, G. 2013. Species distribution modelling and imperfect detection: comparing occupancy versus consensus methods. – Diversity and Distributions 19: 996-1007.

Cutler, D.R. et al. 2007. Random forests for classification in ecology. – Ecology 88: 2783–2792.

Drake, J.M. et al. 2006. Modelling ecological niches with support vector machines. – Journal of Applied Ecology 43: 424–432.

Dudík, M. et al. 2005. Correcting sample selection bias in maximum entropy density estimation. – Advances in Neural Information Processing Systems. Vol. 18.

Dupin, M. et al. 2011. Effects of the training dataset characteristics on the performance of nine species distribution models: application to *Diabrotica virgifera virgifera*. – *PLoS ONE* **6**: e20957. doi:10.1371/journal.pone.0020957.

Elith, J. and Graham, C. 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. – Ecography 32: 66–77.

Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.

Engler, R. et al. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. – Journal of Applied Ecology 41: 263–274.

Ferrier, S. et al. 2002. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. – Biodiversity and Conservation 11: 2275–2307.

Franklin, J. et al. 2009. Effect of species rarity on the accuracy of species distribution models for reptiles and amphibians in southern California. – Diversity and Distributions 15: 167–177.

Gibson, L. et al. 2007. Dealing with uncertain absences in habitat modelling: a case study of a rare ground-dwelling parrot. – Diversity and Distributions 13: 704–713.

Gottschalk, T. K. et al. 2011. Influence of grain size on species–habitat models. – *Ecological Modelling* 222: 3403–3412.

Graham, C.H. et al. 2004. New developments in museum-based informatics and applications in biodiversity analysis. – Trends in Ecology and Evolution 19: 497–503.

Grenouillet, G. et al. 2011. Ensemble modelling of species distribution: the effects of geographical and environmental ranges. – Ecography 34: 9-17.

Guisan, A. and Zimmermann, N.E. 2000. Predictive habitat distribution models in ecology. – Ecological Modelling 135: 147–186.

Guisan, A. et al. 2007. Sensitivity of predictive species distribution models to change in grain size. – Diversity and Distributions 13: 332–340.

Guo, Q. et al. 2005. Support vector machines for predicting distribution of Sudden. Oak Death in California. – Ecological Modelling 182: 75–90.

Hanberry, B.B. et al. 2012. Sample sizes and model comparison metrics for species distribution models. – Ecological Modelling 227: 29–33.

Hernandez, P.A. et al. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. – Ecography 29: 773–785.

Hernandez, P.A. et al. 2008. Predicting species distributions in poorly-studied landscapes. – Biodiversity and Conservation 17: 1353–1366.

Hijmans, R. J. and Elith, J. 2013. dismo: species distribution modeling with R. R project, species distribution modelling with R. http://cran.ms.unimelb.edu.au/web/packages/dismo/vignettes/sdm.pdf.

Hirzel, A.H. et al. 2002. Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? – Ecology 83: 2027–2036.

Jiménez-Valverde, A. et al. 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. – Community Ecology 10: 196–205.

Kadmon, R. et al. 2003. A systematic analysis of factors affecting the performance of climatic envelope models. – Ecological Applications 14: 401–413.

Li, J. and Hilbert, D.W. 2008. LIVES: a new habitat modelling technique for predicting the distribution of species' occurrences using presence-only data based on limiting factor theory. – Biodiversity and Conservation 17: 3079–3095.

Liu, C. et al. 2005. Selecting thresholds of occurrence in the prediction of species distributions. – Ecography 28: 385–393.

Liu, C. et al. 2011. Measuring and comparing the accuracy of species distribution models with presence-absence data. – Ecography 34: 232–243.

Liu, C. et al. 2013a. Selecting thresholds for the prediction of species occurrence with presence-only data. – J. Biogeogr. 40: 478-489.

Liu, C. et al. 2013b. Species distribution modelling for conservation planning in Victoria, Australia. – Ecological Modelling 249: 68-74.

Liu, C. et al. 2016. On the selection of thresholds for predicting species occurrence with presence-only data. – Ecology and Evolution 6: 337–348.

Liu, C. et al. 2018. The first six principal components derived from eighteen environmental variables. – Mendeley Data (v1), <http://dx.doi.org/10.17632/kyy8f8w79j.1>.

Lobo, J.M. and Tognelli, M.F. 2011. Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. – Journal for Nature Conservation 19: 1–7.

Loe, L.L. et al. 2012. Effects of spatial scale and sample size in GPS-based species distribution models: are the best models trivial for red deer management? – European Journal of Wildlife Research 58: 195–203.

Marmion, M. et al. 2009. The performance of state-of-the-art modelling techniques depends on geographical distribution of species. – Ecological Modelling 220: 3512–20.

Mateo, R.G. et al. 2012. Do stacked species distribution models reflect altitudinal diversity patterns? – PLoS ONE 7: e32586. doi:10.1371/journal.pone.0032586.

Meynard, C.N. and Kaplan, D.M. 2013. Using virtual species to study species distributions and model performance. – Journal of Biogeography 40: 1–8.

McPherson, J.M. et al. 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? – Journal of Applied Ecology 41: 811–823.

Naimi, B. and Araujo, M.B. 2016. sdm: a reproducible and extensible R platform for species distribution modelling. – Ecography 39: 368-375.

Pearce, J. and Ferrier, S. 2000. An evaluation of alternative algorithms for fitting species distribution models using logistic regression. – Ecological Modeling 128: 127–147.

Pearson, R. G. et al. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. – Journal of Biogeography 34: 102–117.

Phillips, S. J. et al. 2006. Maximum entropy modeling of species geographic distributions. – Ecological Modelling 190: 231–259.

Phillips, S. J. and Dudík, M. 2008. Modeling of species distributions with MaxEnt: new extensions and a comprehensive evaluation. – Ecography 31: 161–175.

Phillips, S. J. et al. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. – Ecological Applications 19: 181–197.

Pinheiro, J. et al. 2015. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-122. <URL: https://CRAN.R-project.org/package=nlme>.

Pittman, S.J. and Brown, K.A. 2011. Multi-Scale Approach for Predicting Fish Species Distributions across Coral Reef Seascapes. – PLoS ONE 6: e20583. doi:10.1371/journal.pone.0020583.

Qiao, H. et al. 2015. No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation. – Methods in Ecology and Evolution 6:1126-1136.

Reese, G.C. et al. 2005. Factors affecting species distribution predictions: a simulation modelling experiment. – Ecological Applications 15: 554–564.

Robertson, M.P. et al. 2001. A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. – Diversity and Distribution **7**: 15–27.

Santika, T. and Hutchinson, M.F. 2009. The effect of species response form on species distribution model prediction and inference. – Ecological Modelling 220: 2365–2379.

Segurado, P. and Araújo, M.B. 2004. An evaluation of methods for modelling species distributions. – Journal of Biogeography 31: 1555–68.

Seo, C. et al. 2009. Scale effects in species distribution models: implications for conservation planning under climate change. – Biology Letters 5: 39–43

Stockwell, D.R.B. and Peters, D.P. 1999. The GARP modelling system: problems and solutions to automated spatial prediction. – International Journal of Geographic Information Systems 13: 143–158.

Stockwell, D.R.B. and Peterson, A.T. 2002. Effects of sample size on accuracy of species distribution models. – Ecological Modelling 148: 1–13.

Thuiller, W. et al. 2003. Generalized models vs. classification tree analysis: Predicting spatial distributions of plant species at different scales. – Journal of Vegetation Science 14: 669–680.

Thuiller, W. 2004. Patterns and uncertainties of species' range shifts under climate change. – Global Change Biol. 10: 2020–2027.

Thuiller, W. et al. 2009. BIOMOD - a platform for ensemble forecasting of species distributions. – Ecography 32: 369–373.

Warton, D. I. and Shepherd, L. C. 2010. Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. – Annals of Applied Statistics 4: 1383−1402.

White, J.W. et al. 2014. Ecologists should not use statistical significance tests to interpret simulation model results. – Oikos 123: 385−388.

Wisz, M.S. and Guisan, A. 2009. Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. – BMC Ecology 9: 8, doi:10.1186/1472-6785-9-8.

Wisz, M.S. et al. 2008. **Effects of sample size on the performance of species distribution models.** – Diversity and Distributions 14: 763–773.

Supplementary material Appendix 1–4.

**Figure Legends**

Figure 1. Response of model accuracy (the area under the receiver's operating characteristic curve – AUC, median of 200 replicates) to the number of random points (NRP) used in model training for three levels of the number of training presences (NTP: 20, 160 and 640 with solid, dashed and dotted lines respectively) for three levels of prevalence (0.05, 0.25 and 0.75) and for five modelling techniques (artificial neural network – ANN, boosted regression tree – BRT, generalized additive model – GAM, generalized linear model – GLM and support vector machine – SVM) with two weight settings (UTW: unequal total weight, and ETW: equal total weight). The ten levels (1 to 10) of NRP correspond to 10, 20, 40, 80, 160, 320, 640, 1280, 2560 and 5120 respectively.

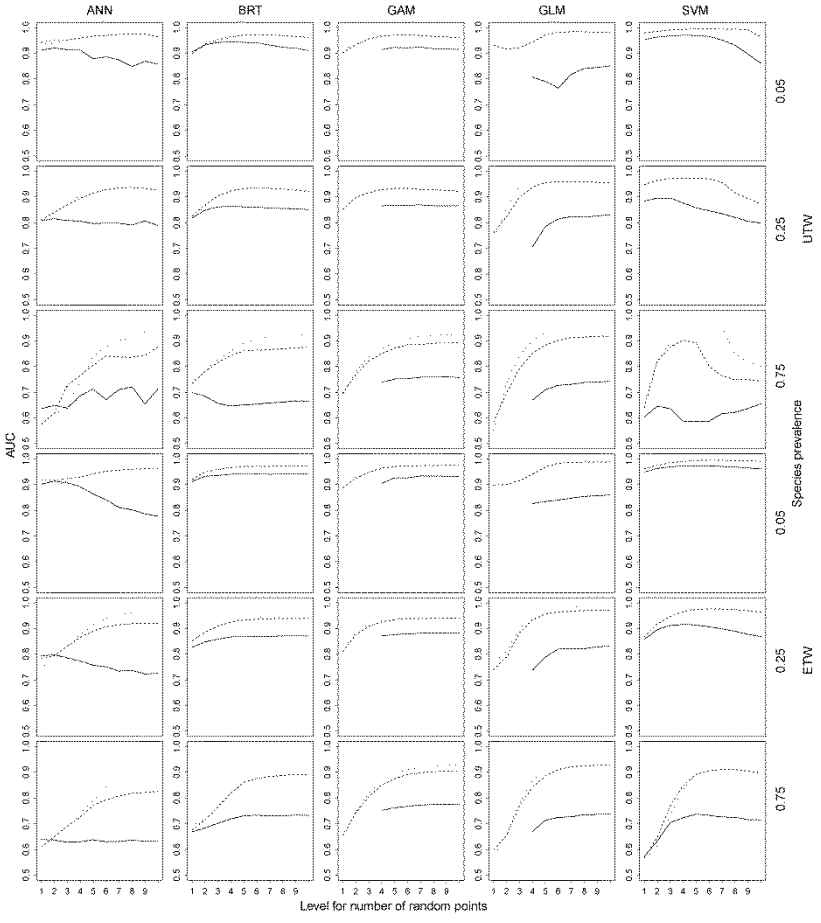

Figure 2. Response of model accuracy (the area under the receiver's operating characteristic curve – AUC, median of 200 replicates) to the number of random points (NRP) used in model training for three levels of the number of training presences (NTP: 20, 160 and 640 with solid, dashed and dotted lines respectively) for three levels of prevalence (0.05, 0.25 and 0.75) and for five modelling techniques (multivariate adaptive regression splines – MARS, mixture discriminant analysis – MDA, Maxent, random forest – RF and ecological niche factor analysis – ENFA) (with unequal total weight setting). The ten levels (1 to 10) of NRP correspond to 10, 20, 40, 80, 160, 320, 640, 1280, 2560 and 5120 respectively.

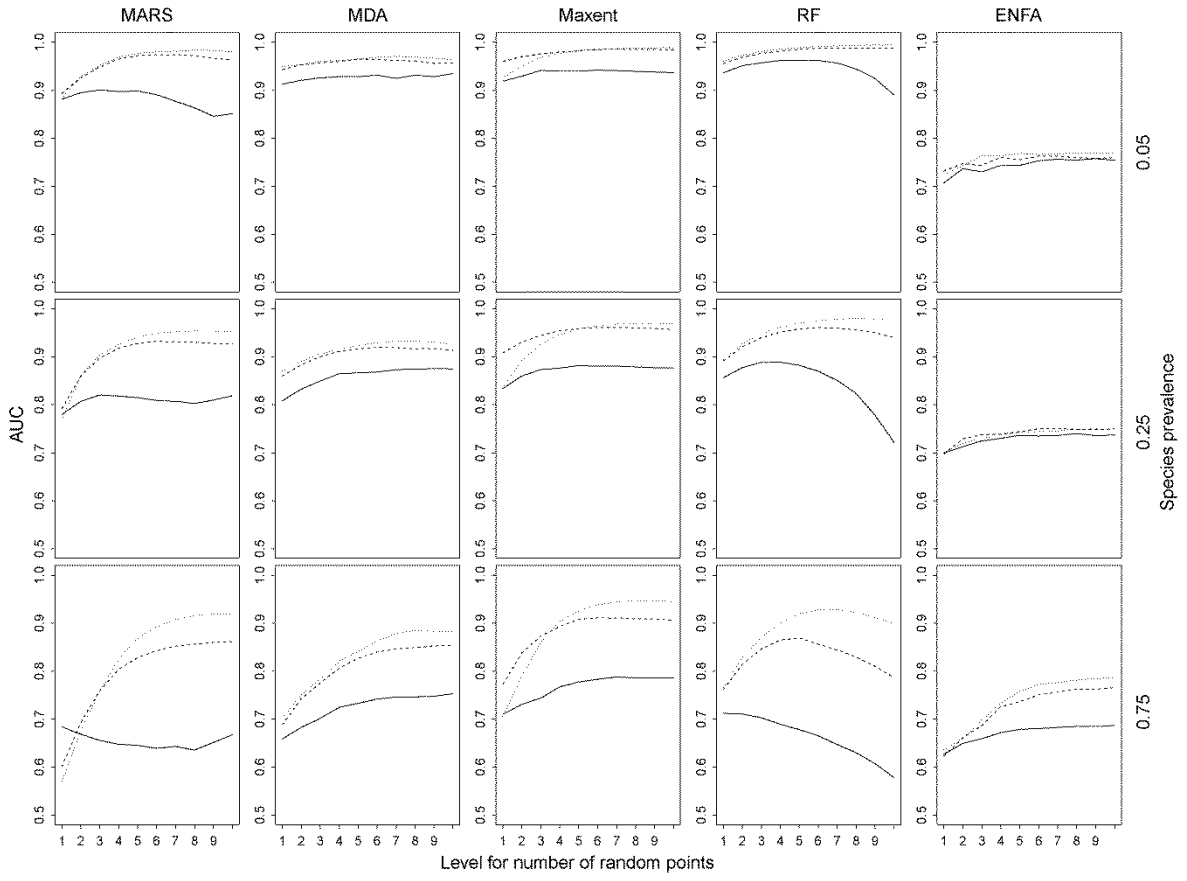

Figure 3. Examples of the four identified patterns for the response of model accuracy (here measured as the area under the receiver's operating characteristic curve – AUC) to the number of random points (NRP) used in model training: (B) Increasing, (B) Decreasing, (C) Arch-shaped, and (D) Horizontal. The solid lines are the medians, the dashed lines are the 2.5 and 97.5$^{th}$ percentiles, and dotted lines are for model predictions. The ten levels (1 to 10) of NRP correspond to 10, 20, 40, 80, 160, 320, 640, 1280, 2560 and 5120 respectively. See Supplementary material Appendix 2 Fig. A22 for details.

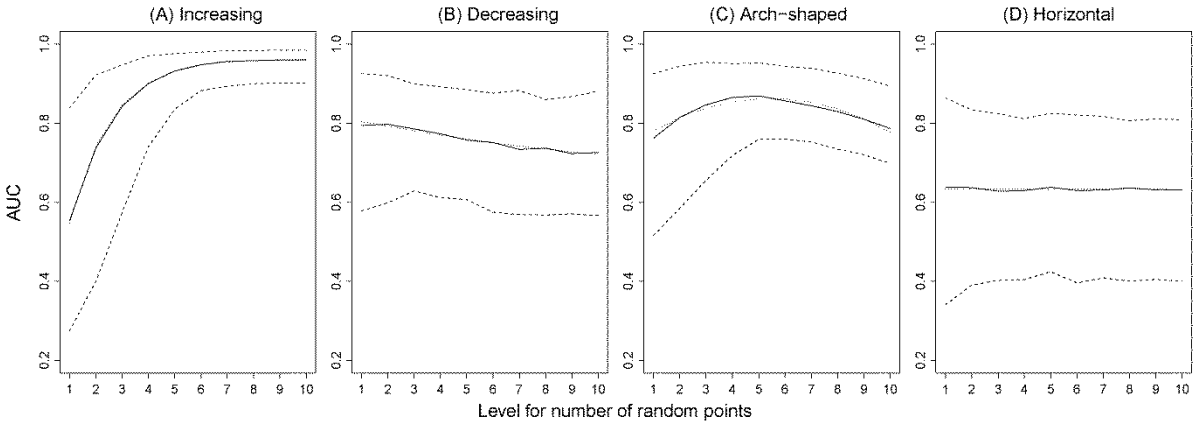

Figure 4. Response of model accuracy (the area under the receiver's operating characteristic curve – AUC, median of 200 replicates) to the number of training presences (NTP) for two strategies (small multiplier and large number) for the number of random points used in model training (solid and dashed lines for small multiplier and large number strategies with UTW respectively, and dotted and dash-dotted lines for small multiplier and large number strategies with ETW respectively) for three levels of prevalence (0.05, 0.25 and 0.75) and for five modelling techniques (artificial neural network – ANN, boosted regression tree – BRT, generalized additive model – GAM, generalized linear model – GLM and support vector machine – SVM) in the upper three rows of panels and another five modelling techniques (multivariate adaptive regression splines – MARS, mixture discriminant analysis – MDA, Maxent, random forest – RF and ecological niche factor analysis – ENFA) in the lower three rows of panels. The nine levels (1 to 9) of the number of training presences correspond to 10, 20, 40, 80, 160, 320, 640, 1280 and 2560 respectively.

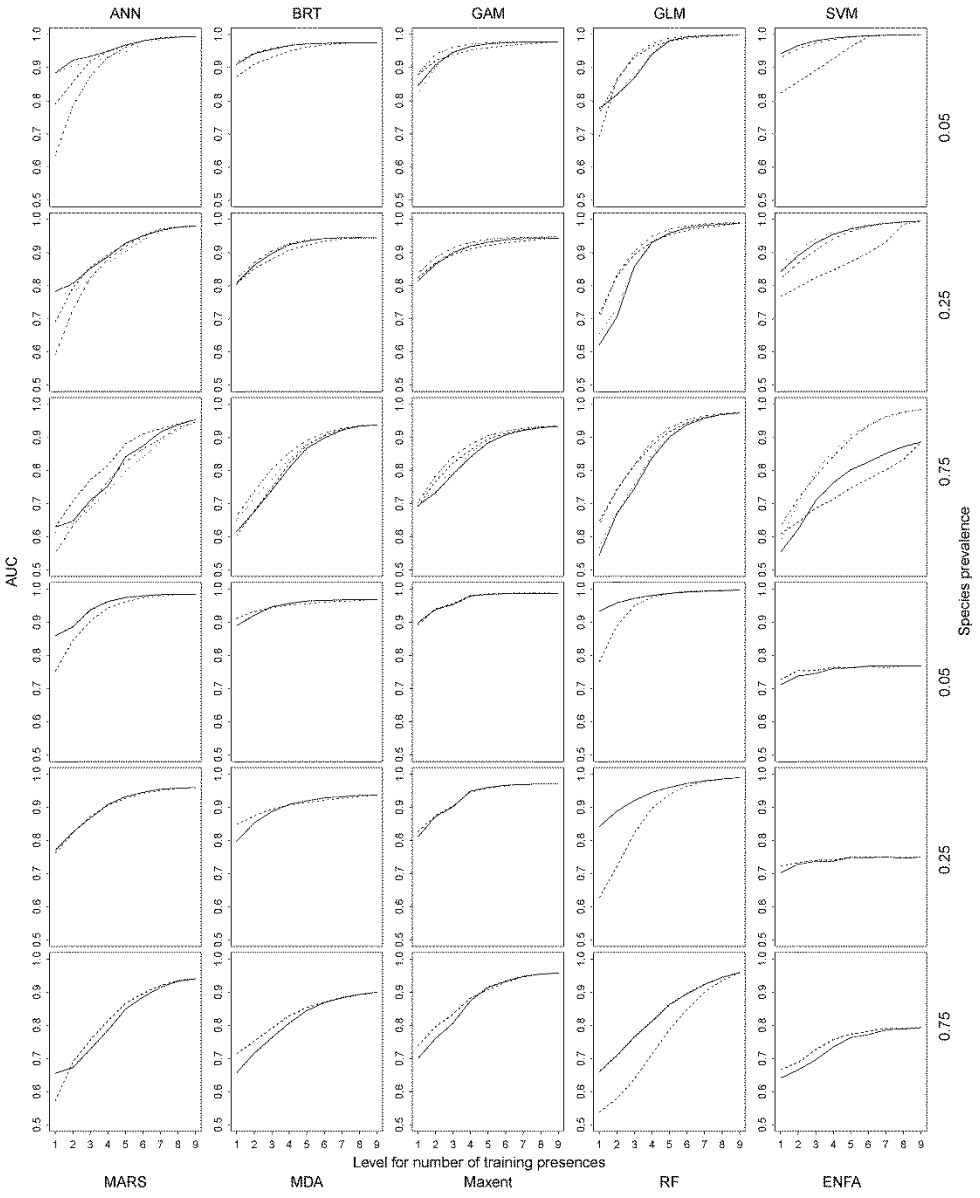

Figure 5. The accuracy (the area under the receiver's operating characteristic curve – AUC) of single and ensemble models (within each segment) for five modelling techniques (artificial neural network – ANN, boosted regression tree – BRT, generalized additive model – GAM, generalized linear model – GLM and support vector machine – SVM) under the combinations of the following parameters: prevalence (0.05, 0.25 and 0.75), number of training presences (NTP: 20, 160 and 640), strategy for number of random points (NRP: S for small multiplier strategy and L for large number strategy), and weight setting (U for unequal total weight and E for equal total weight). Solid symbols represent median values and open symbols represent the 2.5 and 97.5th percentiles of AUC values. Triangles and circles represent single models and ensemble models respectively.

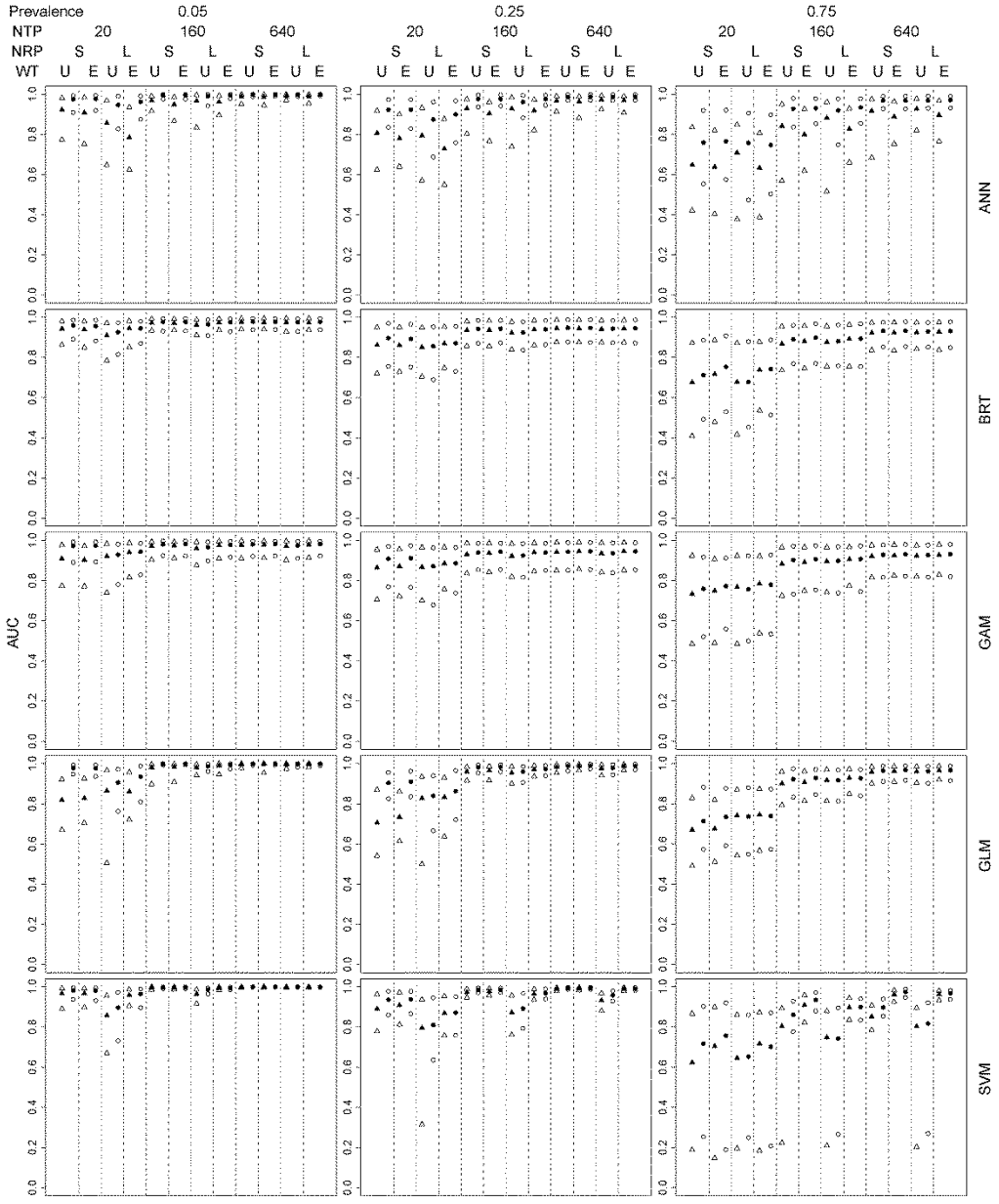

Figure 6. The accuracy (the area under the receiver's operating characteristic curve – AUC) of single and ensemble models (within each segment) for five modelling techniques (multivariate adaptive regression splines – MARS, mixture discriminant analysis – MDA, Maxent, random forest – RF and ecological niche factor analysis – ENFA) under the combinations of the following parameters: prevalence (0.05, 0.25 and 0.75), number of training presences (NTP: 20, 160 and 640), and strategy for number of random points (NRP: S for small multiplier strategy and L for large number strategy). Solid symbols represent median values and open symbols represent the 2.5 and 97.5th percentiles of AUC values. Triangles and circles represent single models and ensemble models respectively.

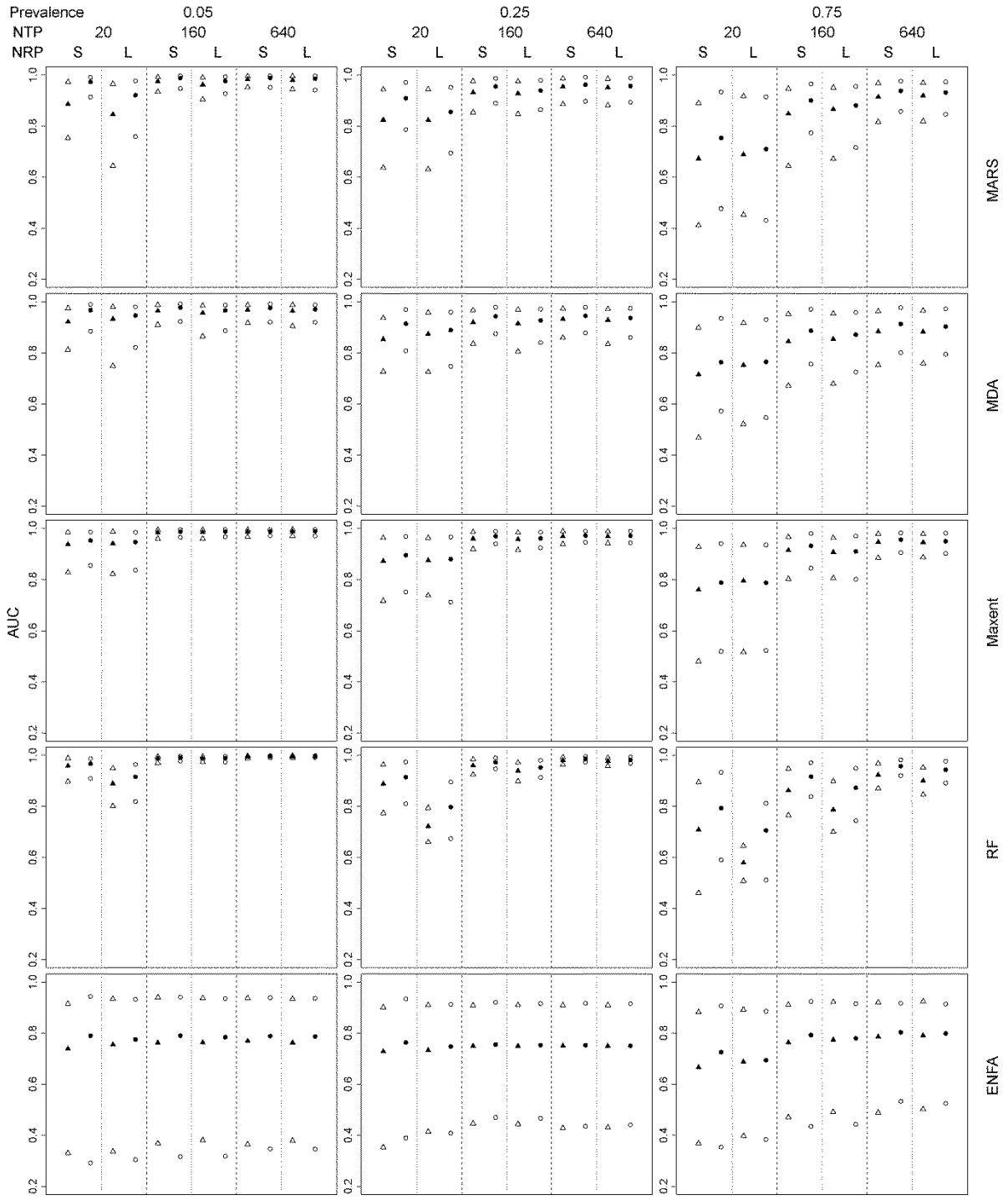

**Table Legend**

Table 1 Preferable modelling strategy for the ten modelling techniques (artificial neural network – ANN, boosted regression tree – BRT, generalized additive model – GAM, generalized linear model – GLM, support vector machine – SVM, multivariate adaptive regression splines – MARS, mixture discriminant analysis – MDA, Maxent, random forest – RF and ecological niche factor analysis – ENFA) according to the number of random points (NRP) used as pseudo-absences or background sites and weight setting in terms of single model versus ensemble model, rare species versus common and very common species, and small number of training presences (NTP) versus large NTP. S_r: small multiplier strategy for NRP with a NRP/NTP ratio r; L: large number strategy for NRP. Since equal total weight is better than unequal total weight for all the scenarios except single ANN models with small NTP, only the preferred strategy for NRP is shown in this table. Since both small multiplier strategy versus large number strategy for NRP and single model versus ensemble model have no obvious effect on the modelling result for ENFA, it is not shown in this table. When using ENFA, we suggest just using the single model with the large number strategy for NRP. For all the other nine techniques, ensemble models are generally better than single models. The bold letters indicate our highly recommended choices, which include equal total weight setting for ensemble SVM models in addition to some situations with small multiplier strategy for NRP. This table can be interpreted in the following way. For example, if a single ANN model is to be built for rare species (rare here refers to species with a low prevalence), when a small number of training presences are available, the best strategy is to use twice as many random points as the training presences and the unequal total weight setting; however, if a large number of training presences are available, the best strategy is to use a large number of random points and the equal total weight setting.

| Conditions | ANN | BRT | GAM | GLM | SVM | MARS | MDA | Maxent | RF |
|---|---|---|---|---|---|---|---|---|---|
| Single model | | | | | | | | | |
| Rare species | | | | | | | | | |
| Small NTP | **S**_2 | S_2 | L | L | **S**_4 | S_2 | S_2 | L | **S**_2 |
| Large NTP | L | L | L | L | L | S_2 | S_2 | L | S_2 |
| Common and very common species | | | | | | | | | |
| Small NTP | S_1 | L | L | L | S_4 | L | L | L | **S**_2 |
| Large NTP | L | L | L | L | S_2 | L | L | L | **S**_2 |
| Ensemble model | | | | | | | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Rare species | | | | | | | | | |
| Small NTP | **S**_2 | S_2 | **S**_16 | **S**_16 | **S**_4 | **S**_2 | **S**_2 | S_2 | **S**_2 |
| Large NTP | S_2 | S_2 | S_2 | S_2 | S_2 | S_2 | S_2 | S_2 | S_2 |
| Common and very common species | | | | | | | | | |
| Small NTP | **S**_2 | S_2 | S_16 | **S**_16 | S_4 | **S**_2 | S_2 | S_2 | **S**_2 |
| Large NTP | S_2 | S_2 | S_2 | S_2 | S_2 | S_2 | S_2 | S_2 | **S**_2 |