Title: Quality of presence data determines species distribution model performance: a novel index to evaluate data quality

Running title: A novel index to evaluate data quality

Authors: Songlin Fei* and Feng Yu

Department of Forestry and Natural Resources, Purdue University

*Correspondence: Songlin Fei, PFEN221E, Purdue University, 715 W. State St., West Lafayette, IN, USA, 47907. Email: sfei@purdue.edu

Date of manuscript draft: August 18, 2015

Manuscript Word Count: 5138

Number of Figures: 5

ABSTRACT

Context Species distribution models (SDMs) are widely used to estimate species' potential distribution at landscape to regional scales. However, the quality of occurrence data is often compromised by sampling bias. The negligence of data quality assessment could raise serious concerns on model accuracy.

Objectives. We propose a model-independent composite measure - representativeness and completeness (RAC) index - to evaluate the quality of species occurrence data. We demonstrate (1) the impact of spatial data quality as measured by RAC on model performance and (2) the feasibility of applying RAC in actual modeling process.

Methods. By using a set of computational experiments on a virtual species, we calculated RAC values for a set of occurrence data (35 runs x 50 datasets x 5 series) representing different degrees of sampling biases. We evaluated model performance (reliability and accuracy) using three different model algorithms; and we associated model performance with RAC values. Two case studies were also conducted to demonstrate the association between RAC and model performance.

Results. Model reliability stabilizes when RAC reaches a threshold of 0.4. Model accuracy also stabilizes when RAC reaches 0.4 or 0.5 for models with or without complete predictors, respectively. Model performance is more sensitive (i.e., has larger variability) to data completeness than representativeness. Our case studies further demonstrated that RAC value is closely related to model performance.

Conclusions. Performance of SDMs is closely related to the quality of species occurrence data, which can be measured by our model-independent composite index - RAC. We recommend a minimum RAC value of 0.4 for reliable and accurate SDM predictions. To improve prediction

accuracy, sampling with multiple centers in a systematic fashion across the environmental space is desired.

Keywords Species distribution modeling, data quality, representativeness, completeness

INTRODUCTION

Species distribution models (SDMs) estimate the relationship between species occurrence or abundance with the corresponding environmental conditions using a set of statistical methods (Elith and Leathwick 2009). Due to the wide availability of species occurrence data and efficient modeling tools, SDMs have received increasing attention from forecasting risk of biological invasions and impacts of climate change to spatial conservation planning and historical biogeography, with thousands of papers being published annually (Franklin 2013). The products of SDMs, usually the predicted distribution maps or habitat suitability maps at landscape to regional-scales, often serve as the foundation and possibly the only reliable information for conservation planning, risk assessment, or resource management implementations (Franklin 2009).

However, the quality of species distribution data is often compromised by sampling bias in terms of spatial distribution, especially for presence-only data. For example, sampling intensity may not be consistent across regions. Some regions are under-sampled due to poor site accessibility or incomprehensive survey plan, while other regions are over-sampled (Peterson and Holt 2003; Phillips et al. 2006). In addition, sometime only part of the presence data was used in the modeling process. Therefore, the accuracy and reliability of SDMs are likely compromised by incomplete or biased survey data.

Paradoxically, the lack of comprehensive observations is probably the most important reason for using SDMs to extend the data availability for conservation planning and resource management purposes, predicting species distributions in the remote areas that field surveys are not able to cover. To improve the performance of SDMs based on ‘imperfect’ species distribution data, a set of methodological approaches have been advanced in the recent years, such as sampling bias correction (Kramer-Schadt et al. 2013), data quality discrimination (Hortal et al. 2001), appropriate predictor selection and model parameterization (Elith and Leathwick 2009; Merow et al. 2013), and consensus forecasting (Araújo and New 2007). However, inherent limitations of the sampling data cannot be fully corrected only through these approaches (Merow et al. 2013; Phillips and Elith 2013).

This raises an important question addressed in this paper: can we estimate model performance, both in reliability and accuracy, based on the quality of species occurrence data prior to the application of SDMs? Among many components of SDMs, species occurrence data are the foundation for the prediction of species occurrence probability or habitat suitability. Therefore, the quality of the occurrence data can be a decisive contributor for the performance of SDMs. Currently, limited systematic frameworks have been suggested to evaluate the quality of species distribution data used for SDMs (but see Hortal and Lobo 2005; Lobo and Martín-Piera 2002; Luoto et al. 2005; Reese et al. 2005), and no quantitative tool is available to evaluate the quality of species distribution data.

The key questions in evaluating the quality of occurrence data are: what are the most important factors that influence the quality of species distribution data, and how can the quality of species distribution data be quantitatively assessed? Fortunately, all existing SDMs share one important characteristic -- they are essentially all statistical models (Austin 2007; Elith et al. 2006). These statistical models require two underlying assumptions to comply with the basis of ecological niche theory: (1) the sampling occurrence should be representative of the extent of the ecological niche in the environmental space (whether it is in a fundamental or realized niche depends on the model method) and (2) the sampling occurrence should be at equilibrium, covering the entire extent of the ecological niche in the environmental space (Araújo and Pearson 2005; Guisan and Thuiller 2005). Representativeness is defined as the sampling data being randomly distributed throughout the ecological niche; while equilibrium is defined as the occurrence data comprehensively covering the entire niche and being absent in the locations outside of the niche, which we refer to as completeness (Araújo and Pearson 2005; Guisan and Thuiller 2005; Václavík and Meentemeyer 2012). Although the two assumptions are rarely satisfied in actual survey data (Acevedo et al. 2012), they can be used to quantify the quality of species distribution data, especially presence-only data. More specifically, we can calculate the degree of representativeness and completeness of the data in the environmental space of known niche or observed occurrence range.

In this study, we proposed a novel quantitative measure, representativeness and completeness (RAC) index, to assess SDMs' predictive reliability and accuracy based on the two assumptions mentioned above by focusing only on the characteristics of species occurrence data. The application of RAC will allow modelers to determine the quality of occurrence data and the confidence of modeling results independent of SDMs. We studied the relationship between the sampling pattern of occurrence data in the environmental space (constructed by bioclimatic variables) and model performance. Our hypothesis is that the quality of occurrence data, as measured by our RAC index, is positively related to model performance. By focusing on the occurrence data, modelers can discern a very basic but essential assessment of the 'usefulness' of their data and the expected accuracy and reliability. We expect this study will greatly benefit the research and practices of the macrosystem ecology, biogeography, and conservation community by avoiding seriously biased or incorrect model outputs that are based on species distribution data of unsatisfactory quality.

METHODS

We examined the influence of occurrence data quality on model performance by using a virtual species living in a scaled environmental space with the following major steps. First, we created a gradient of occurrence patterns representing different degrees of sampling biases, and measured the corresponding data quality with RAC. Second, we evaluated model performance (reliability and accuracy) for three different model algorithms using the 'binned' evaluation method for each datasets. Third, the relationship between data quality and model performance was analyzed and critical thresholds of RAC were identified where the model performance was significantly improved and stabilized.

Virtual species in a scaled environmental space

We created a virtual species living in a scaled environmental space. Two environmental variables (predictors): temperature (*T*) and precipitation (*P*), both scaled with a range of 0 to 1, were used to create the environmental space (akin to Phillips and Elith 2010). The extent of this environmental space includes all existing environmental conditions that occurred in all known or observed areas, which is equivalent to the 'biotope' according to the classic niche concept of Hutchinson (Franklin 2009). We assumed that the true probability of presence of this virtual species is defined as $p_T = (T+P)/2$ in this environmental space. The utilization of this scaled environmental space allows our findings to be applicable to both generalist and specialist species since the niche studied here is in a relative scale (0 to 1) regardless of its actual niche width.

Point pattern gradients

We designed five data series to represent possible spatial distribution patterns of sample locations. To clarify, the simulated point patterns in the following procedure are not the representations of the underlying species distribution, which could have different patterns (e.g., clustered, random, etc.) that are often species or taxonomic groups specific; rather, the simulated patterns are the representations of possible sampling results due to different degree of sampling biases.

Each of the five data series has 50 unique point patterns that evolve gradually from highly clustered to completely random following the method by Ong et al. (2012) (**Figure 1**). In order to obtain a continuous and gradual transition, each point pattern was created by combining point samples from a random distribution in one or several grid squares (Z1) and the complimentary subareas (Z2, which is the total area minus Z1) of the environmental space (Ong et al. 2012). Because sample size could be an important factor influencing model performance where small sample size (n<30) cannot produce consistent predictions (Wisz et al. 2008), we used a total of 1,000 points to generate each point pattern to ensure sample size will not impact our analysis. Z1 contains 990 point locations and Z2 contains 10 point locations (see **Appendix 1** for detailed procedure for point pattern generation). We used the ‘Create Random Points’ tool in ArcMap (ESRI, Redlands, CA, USA) to generate the random point patterns in Z1 and Z2 with a Poisson distribution.

Although each dataset was generated through combining random points from Z1 and Z2, our preliminary analysis indicated that it did not completely eliminate the possibility of regional pattern bias in the environmental space (e.g., highly clustered in one subarea of Z1). Therefore, to reduce the possible spatial bias due to random chances, we ran 7 iterations to generate different sets of random points (i.e., each iteration generates different point patterns representing locations of presence/pseudo-absence of the virtual species for each of the 50 datasets in each series). ‘Presence’ or ‘absence’ of each point was determined via Bernoulli trials (0 and 1) based on the occurrence probability (P_T) at a given location. Note that the term ‘absence’ was not intended to represent the locations where the environmental conditions were not suitable for this virtual species; rather, it was used to simulate modeling procedure that based only on presence data (Franklin 2009; Phillips et al. 2009).

To simulate the presence/absence at a given point due to random chances (i.e., presence of a given species may not be observed at a given location during one sampling effort due to random effects), we ran 5 replications (5 sets of Bernoulli trials) to assign presence/absence values for each of the 7 iterations. Therefore, each of the 50 datasets in each series has a total of 35 runs of occurrence patterns, allowing sufficient statistical power given our computational capacity (5 series x 50 datasets/series x 7 iterations/dataset x 5 runs/iteration = 8,750 runs). Data analyses were conducted on each of the 8,750 runs.

Measurement of species distribution data quality

We created a composite measure - RAC (representativeness and completeness) index - to quantify the quality of species distribution data. RAC includes the calculation of nearest neighbor statistics (a measure of representativeness) and completeness ratio (a measure of completeness). The completeness ratio is the degree of coverage of the observed species occurrence in the corresponding environmental space of its known or observed niche. For a two-dimensional environmental space, the completeness ratio (Ω) can be calculated by multiplication of the completeness of each dimension as follows:

$$\Omega = \prod_{i}^{2} \frac{C_i}{N_i}$$

where C_i is the observed occurrence range of the environmental variable i, and N_i is the corresponding known or observed niche described by the environmental variable i. The range of Ω varies from 0 to 1, which reflects the degree of completeness (Kadmon et al. 2003).

We used nearest neighbor statistic (R) to measure representativeness based only on occurrence localities (presence points). The core concept of nearest neighbor analysis is to calculate the mean Euclidean distances between each paired nearest neighbors, and then compare the result with the expected or theoretical situation. According to Clark and Evans (1954), the mean of observed distance D_o can be calculated as:

$$D_o = \frac{\sum_{i=1}^{n} r_i}{N}$$

where r_i is the distance for the ith point to its paired nearest neighbor and N is the total number of points in the environmental space. The expected mean nearest neighbor distance D_e in a uniform random distribution can be calculated as:

$$D_e = \frac{1}{2}\sqrt{\frac{A}{n}}$$

where A is the total area of the environmental space in the known or observed niche and n is the number of points. The reason we used uniform random distribution to calculate D_e is that the critical assumption required for presence-only data based SDMs is the random sampling of space for accumulating presence only observations (Royle et al. 2012). Nearest neighbor statistic (R) is the ratio of observed (D_o) to expected (D_e) mean nearest neighbor distance ($R = D_o/D_e$). R approaches to 0 if the point pattern is completely clustered, and R approaches to 1 if the point pattern is uniform random (Clark and Evans 1954, see **Appendix 2** for examples). Our composite measure (RAC) of the spatial pattern of the species distributions localities was defined as the product of completeness (Ω) and the nearest neighbor statistic (R),

$$RAC = R \times \Omega$$

RAC varies from 0 (total deviation from completeness and representativeness) to 1 (perfect completeness and representativeness).

Simulation of SDMs

We designed three model algorithms to simulate different possible scenarios of actual modeling methods in terms of the selection of environmental variables akin to Phillips and Elith (2010). Model 1, $p_1 = 0.25 + 0.5 \times T$, estimates the probability of species occurrence only based on temperature. In practice, this could result from some of the essential predictors controlling the species distribution not being correctly identified or some predictors being mistakenly excluded due to a certain SDM method (Phillips and Elith 2010). Model 2, $p_2 = 0.25 \times (T + P) + 0.125 \times (T + P)^2$, estimates the probability of presence from both of the variables (T and P) as a comprehensive suite of the possible combinations of the environmental variables including interaction terms. Model 3, $p_3 = (T+P)/2$, is the ‘true’ probability of presence of the species.

Model evaluation

Model performance is generally described as predictive accuracy, which has two components: discrimination and calibration (Vaughan and Ormerod 2005). Discrimination is the ability of the model to correctly distinguish occupied from unoccupied sites, whereas calibration measures the agreement between predicted probabilities of occurrence and observed proportions of sites occupied (Pearce and Ferrier 2000). In this study, we focused only on calibration - the numerical accuracy of prediction (Phillips and Elith 2010; Vaughan and Ormerod 2005). Discrimination was not calculated because there is no true absence in our environmental space.

Calibration can be examined graphically via a ‘binned’ method (Pearce and Ferrier 2000; Vaughan and Ormerod 2005) by plotting the median values of predicted probabilities in each of the predefined predicted probability intervals against the fraction of the actual observed localities (marked as ‘presence’) vs. the total localities within each of these probability intervals (see example in Appendix 3). The overall calibration across all the predicted probability intervals, which indicates the general model performance, can be obtained from the slops of a linear regression line of these plots. In general, slope = 1 indicates good model calibration, slope > 1 indicates underestimation, and slope <1 indicates overestimation (Vaughan and Ormerod 2005). In this study, we divided the predicted probability into 10 intervals and plotted the median in each interval against the fraction of the actual observed localities.

Association between model performance and data quality

To analyze the association between model performance and RAC, we plotted the mean and standard error of the slopes from the linear regression of the binned method against the mean

RAC value of the 35 runs in each dataset for all three models (**Figure 2**). Mean of the slope was used to measure the accuracy (average slope approaching 1 indicates high accuracy); while standard error of the slope was used to measure model reliability (standard error is minimized for all 35 runs when models are reliable). In addition, we used multivariate adaptive regression splines (MARS) to further illustrate the relationships between RAC values and model reliability and accuracy, and to detect critical RAC thresholds where model performance are significantly improved.

RESULTS

In general, we found that both model reliability and accuracy increase with data quality as measured by RAC (**Figure 2**). As point pattern evolves from highly clustered to near uniform random, model performance improves. Note that the trend does not proceed gradually along the point pattern gradient. Instead, model performance, in terms of both model reliability (**Figure 2a**) and accuracy (**Figure 2b**), 'jumps' or 'turns' significantly at certain critical RAC values. These critical points serve as important indicators of model performance for different model algorithms and point patterns.

Model reliability dramatically improves as RAC value approaches to 0.4, and stabilizes after this threshold regardless of model algorithms or spatial point patterns (**Figures 2a and 3a**). Model accuracy also follows the same trend. Model accuracy improves as RAC value approaches to the thresholds [0.40 for models with completed environmental variables (Model 2 and Model 3) and 0.50 for models missing key environmental variables (Model 1)], and stabilizes after that

(**Figures 2b and 3b**). All models, except Model 1 in Series E, have an accuracy of greater than 0.7 after RAC value reaching the above thresholds (**Figure 4**).

Of the two components that defines RAC index, both representativeness and completeness values are positively related to model accuracy; while model performance is more sensitive (i.e., has larger variability) to data completeness than representativeness (**Figure 5**). Model accuracy becomes stable when representativeness value reaches 0.40 for models with completed environmental variables (Model 2 and Model 3) and 0.85 for models missing key environmental variables (Model 1) (**Figure 5a)**. Whereas model accuracy becomes stable only after completeness value reaches 0.95 for all models (**Figure 5b**).

Spatial pattern of presence-only data has a strong influence on model performance. Models in series A have higher accuracy and reliability than any other series (**Figures 2 and 4**). In general, both model accuracy and reliability are high even when RAC value is below the threshold. This is because data in series A have multiple clusters (9 total) evenly-distributed across the environmental space of the virtual species; while data in other series have a single cluster. This again confirms our hypothesis that model performance is very sensitive to the spatial distribution of species occurrence data, and even to how the points are clustered spatially.

Models with complete environmental variables (Model 2 and 3), regardless of whether interactions were included in the model, often achieve model stability easier (with lower RAC values) than models that lack one or more essential environmental variables (Model 1) (**Figure 2**). Model 1 is designed to simulate the possible practical scenario that some important variables are not selected in the modeling process or the model mistakenly excluded some environmental variables. This indicates the importance of predictor selection since models with incomplete environmental predictors require a higher degree of completeness and representativeness from the distribution data to achieve better model performance, which in turn increases the cost and workload of conducting species survey.

CASE STUDIES

To demonstrate the use of RAC to measure the quality of species distribution data and its implications of model performance, we modeled the habitat suitability of two invasive species, *Prosopis farcta* (PRFA) and *Imperata cylindrica* (IMCY). Presence data of the two species were obtained from the Global Biodiversity Information Facility (GBIF, www.gbif.org/), a widely used database for SDMs. Native ranges for these two species were used to define their ecological niches. IMCY was selected to represent the case of poor quality of presence data since the presence data were clustered in a portion of its native range in the environmental space even though it has a large sample size (n= 4,913); while PRFA was selected (n=281) to represent better data quality in terms of spatial coverage.

Unlike the virtual species modeled above in the idealized landscape, where environmental space overlaps with geographical space, we need to convert the environmental variables within the geographical range to environmental space to calculate RAC for actual species. First, we obtained six environmental variables (annual mean temperature, isothermality, mean temperature of coldest quarter, annual mean precipitation, coefficient of variation of monthly precipitation, and precipitation of driest quarter) from WorldClim (http://www.worldclim.org, Hijmans et al.

2005) based on knowledge from variable selections of other studies (Austin 2007; Austin and Smith 1989; Elith and Leathwick 2009). Second, we used principal component analysis (PCA) on these six variables to reduce the environmental dimensions to two orthogonal components. We measured the correlation among the six variables to ensure that the data were suitable for PCA. The results indicated that in the correlation matrix, each variable had at least one correlation coefficient > 0.3 and the overall Kaiser-Meyer-Olkin (KMO) value was >0.5, satisfying the minimum requirement for conducting a PCA (Kaiser 1974). The two major components, PCA1 and PCA2, explained 59.85% and 22.84% of the variability for PRFA and 51.84% and 24.61% for IMCY, respectively. Third, species occurrence data were then re-plotted onto the new two-dimensional environmental space (**Appendix 4**), and their RAC values were calculated using the procedure described in the methods section.

To compare RAC value with other model performance indicators, we used MaxEnt (Version 3.3.3k, Phillips et al. 2006) to model the spatial distribution of PFRA and IMCY based on the same presence data and six environmental variables mentioned above. 'Target Group' background method was applied to reduce the sampling bias, as suggested by Phillips et al. (2009). We used 38 species from our invasive species study (unpublished data) to compose the locations of the target-group background. Area Under Curve (AUC) and True Skill Statistics (TSS) were used to evaluate the model performance (Allouche et al. 2006; Pearce and Ferrier 2000). AUC is a single-value, threshold-independent indicator of the general model performance that is not influenced by species prevalence (Manel et al. 2001; Pearson et al. 2013). However, AUC is not sensitive to the shift of model accuracy and a poorly fitted model may also receive high AUC value (Lobo et al. 2008). On the other hand, although TSS is a threshold-dependent evaluation method, it provides the ability to distinguish between a well fitted model and poorly fitted model, which is also not influenced by the species prevalence (Allouche et al. 2006). We used Maximum Sensitivity plus Specificity (MSS, Jiménez-Valverde and Lobo 2007) to determine the 'presence-absence' threshold to calculate TSS. AUC and TSS are complementary and reflect different perspectives of model performance.

Results from the case studies further illustrated that RAC value is positively associated with model performance. RAC value for PRFA is 0.449 (high degree of completeness, Ω = 0.830 and evenness, R = 0.541). Correspondingly, AUC and TSS for PRFA were 0.991 and 0.902, respectively, indicating good model performance. RAC value for IMCY is 0.215 (lower degree of completeness, Ω = 0.600 and evenness, R = 0.358). Correspondingly, AUC and TSS for IMCY were 0.890 and 0.656, respectively, indicating poor model performance. Larger differences were observed in TSS values than AUC. This is likely due to effect of prevalence of species occurrence (sample size of IMCY was much larger than PRFA), which may result in the falsely higher AUC (Allouche et al. 2006).

DISCUSSION

In this study, we demonstrated that the quality of species distribution data is closely associated with the performance of SDMs. Our results clearly revealed that RAC values are positively correlated with model performance in both reliability and accuracy. In general, model performance stabilizes when RAC reaches 0.40 for models include all necessary environmental variables, regardless of model algorithm used.

The correlation between RAC and predictive accuracy clearly confirmed the conclusion from other studies that sampling bias has a profound impact on predictive accuracy (Kadmon et al. 2003). However, such sampling bias has been difficult to describe in the modeling work without survey information, especially for the presence-only data. In this study, we quantitatively linked sampling bias to the spatial point patterns of occurrence data in the environmental space. Modelers can compare to the proposed thresholds and estimate what the predictive results will be likely achieved when applying these data in their SDMs.

Research implications

Between the two components that determines RAC value, completeness is more sensitive to model performance than representativeness. This partially explains the results that models in series A have higher accuracy and reliability than any other series because data in series A have multiple clusters (9 total) evenly-distributed across the environmental space, resulting a higher completeness and representativeness values compared to data in other series that have a single cluster. Our findings agree with a sampling bias correction study by Fourcade et al. (2014), who found that among five sampling bias correction methods, simple systematic re-sampling (i.e., subsample of records that are regularly distributed) of available data consistently outperformed all other methods across various test conditions.

Our study clearly demonstrated the need of high quality primary data in order to achieve subsequent accurate modeling results. As advocated by many prior studies (e.g., Hijmans 2012; Lobo and Martín-Piera 2002; Reese et al. 2005), we need well-designed surveys to obtain high quality data for better model performance. We suggest the collection or utilization of data with high spatial completeness and representativeness for accurate modeling performance if possible. More specifically, data with multiple centers in a systematic fashion across the environmental space can facilitate the faster achievement of the recommended RAC threshold than mono-centered data.

Research applications

We used a virtual species to demonstrate the necessity of high data quality for accurate and reliable model performance. Although there are some challenges as often encountered in other species modeling efforts, results from our computational experiments can be easily applied in real situations with the following general steps: (1) convert data from geographic space to environmental space; (2) measure RAC value of sample points in the environmental space, and (3) compare the calculated RAC value with the recommended minimum RAC value.

Converting data from geographic space to environmental space can be conducted in two steps. First, environmental variables that are closely related to the distribution of the target species should be selected (e.g., Fei et al. 2012). Second, ordination methods such as principal component analysis (PCA) or environmental niche factor analysis (ENFA, Hirzel et al. 2002) need be used to reduce the environmental dimensions to two orthogonal components. The resulting two orthogonal components can be used to represent the environmental space for the target species. One of the main advantages of calculating RAC values in environmental space is that moving from geographic to environmental space may reduce erroneous, or misleading, “representativeness” values. For example, a species may be clustered in geographic space because it requires specific environmental conditions that are also clustered on the landscape.

Conversion to environmental space should alleviate some of this spatial clustering. The limitation of this approach is that sometimes the resulted primary and secondary components cannot adequately represent the variability of the environmental space.

To measure RAC value in the converted environmental space, we first need to re-plot the occurrence data onto the new two-dimensional environmental space and calculate RAC value using the procedure described in the method section. However, it is not straight-forward to define the boundaries of the environmental space or niche width (realized vs. fundamental niches), a common challenge in species distribution models (Elith and Leathwick 2009; Sax et al. 2013). For our virtual species, we assumed that the fundamental niche, realized niche and 'biotope' are approximately equivalent. This may only apply to broad spatial scales and species have reached distribution equilibrium. In practice, we recommend the use of range maps (an approximation of realized niche) to define the boundaries of the environmental space, as range maps are often available for many species in North America, Europe, and Asia. Digital range maps (or to be digitized from paper versions) can be used to overlay on all environmental variable layers to define environmental spaces. After the calculation of the RAC value for the target species, we can compare it with the threshold value (RAC > 0.4) to determine if the data quality is acceptable.

Cautions need be made when applying the proposed RAC index to assess data quality due to its inherited limitations. Models used to predict species distribution often involve various environmental variables such as climate, soil, terrain, geology, land use, and etc. These variables often have their limitations on thematic, spatial, and temporal scales, and dimension reduction using two principal components may not capture the variations of ecological niches. Both of which can limit the feasibility of the application of the proposed 0.4 RAC threshold. On the other hand, variables used in SDMs are often highly correlated (e.g., Fei et al. 2007; Liang and Fei 2014) and climate factors are the first order variables that influence species ranges limits and are often readily available (Ricklefs and Jenkins 2011; Shen et al. 2012). Therefore, we recommend the application of the proposed RAC index primarily on bioclimatic variables as demonstrated in our case studies. Additional research is needed to study the impact of the inclusion of other non-climatic variables and the interplays among environmental variables on RAC values and the subsequent model performance.

In conclusion, we demonstrated that the performance of SDMs, in terms of reliability and accuracy, is closely related to the quality of species occurrence data. We provided a novel way to evaluate data quality through a composite measure - RAC, which measures the degree of proximity to ideal representativeness and completeness. Modelers can estimate the quality of model results by calculating the RAC values of their occurrence data and comparing them with the recommended critical thresholds (RAC > 0.4) prior to their modeling effort. Unlike other commonly used model evaluation methods, RAC is model-independent. Modelers can use this method to preselect and eliminate unsuitable species distribution data before running any model. A user-friendly tool is needed to easily calculate the RAC value for various species and model efforts for the benefit of the ecology, biogeography, and conservation communities.

ACKNOWLEDGMENT

We thank Drs. Janet Franklin, Jeffrey Dukes, Jane Frankenberger for helpful comments on an earlier versions of the manuscript. We acknowledge funding support from the National Science Foundation (Macrosystems Biology 1241932).

REFERENCES

Acevedo P, Jiménez-Valverde A, Lobo JM, Real R (2012) Delimiting the geographical background in species distribution modelling. J. Biogeogr. 39(8):1383-1390

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

Araújo MB, New M (2007) Ensemble forecasting of species distributions. Trends Ecol. Evol. 22(1):42-47

Araújo MB, Pearson RG (2005) Equilibrium of species’ distributions with climate. Ecography 28(5):693-695

Austin M (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecological Modelling 200(1):1-19

Austin M, Smith T (1989) A new model for the continuum concept. Vegetatio 83(1-2):35-47

Clark PJ, Evans FC (1954) Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology 35(4):445-453

Elith J, Graham CH, Anderson RP et al (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography 29(2):129-151

Elith J, Leathwick JR (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology Evolution and Systematics 40:677-697

Fei S, Liang L, Paillet FL et al (2012) Modelling chestnut biogeography for American chestnut restoration. Divers. Distrib. 18:754-768

Fei S, Schibig J, Vance M (2007) Spatial habitat modeling of American chestnut at Mammoth Cave National Park. For. Ecol. Manag. 252(1-3):201-207

Fourcade Y, Engler JO, Rödder D, Secondi J (2014) Mapping Species Distributions with MAXENT Using a Geographically Biased Sample of Presence Data: A Performance Assessment of Methods for Correcting Sampling Bias. PloS one 9(5):e97122

Franklin J (2009) Mapping species distributions: spatial inference and prediction. Cambridge University Press Cambridge

Franklin J (2013) Species distribution models in conservation biogeography: developments and challenges. Divers. Distrib. 19(10):1217-1223

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

Hijmans RJ (2012) Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. Ecology 93(3):679-688

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25(15):1965-1978

Hirzel AH, Hausser J, Chessel D, Perrin N (2002) Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology 83(7):2027-2036

Hortal J, Lobo J (2005) An ED-based Protocol for Optimal Sampling of Biodiversity. Biodivers. Conserv. 14(12):2913-2947

Hortal J, Lobo J, Martín-piera F (2001) Forecasting insect species richness scores in poorly surveyed territories: the case of the Portuguese dung beetles (Col. Scarabaeinae). Biodivers. Conserv. 10(8):1343-1367

Jiménez-Valverde A, Lobo JM (2007) Threshold criteria for conversion of probability of species presence to either–or presence–absence. Acta Oecologica 31(3):361-369

Kadmon R, Farber O, Danin A (2003) A systematic analysis of factors affecting the performance of climatic envelope models. Ecological Applications 13(3):853-867

Kaiser HF (1974) An index of factorial simplicity. Psychometrika 39(1):31-36

Kramer-Schadt S, Niedballa J, Pilgrim JD et al (2013) The importance of correcting for sampling bias in MaxEnt species distribution models. Divers. Distrib. 19(11):1366-1379

Liang L, Fei S (2014) Divergence of the potential invasion range of emerald ash borer and its host distribution in North America under climate change. Clim. Change 122(4):735-746

Lobo JM, Jiménez-Valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography 17(2):145-151

Lobo JM, Martín-Piera F (2002) Searching for a predictive model for species richness of Iberian dung beetle based on spatial and environmental variables. Conserv. Biol. 16(1):158-173

Luoto M, Pöyry J, Heikkinen R, Saarinen K (2005) Uncertainty of bioclimate envelope models based on the geographical distribution of species. Global Ecology and Biogeography 14(6):575-584

Manel S, Williams HC, Ormerod SJ (2001) Evaluating presence–absence models in ecology: the need to account for prevalence. Journal of Applied Ecology 38(5):921-931

Merow C, Smith MJ, Silander JA (2013) A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography 36:1058-1069

Ong MS, Kuang YC, Ooi MP-L (2012) Statistical measures of two dimensional point set uniformity. Computational Statistics & Data Analysis 56(6):2159-2181

Pearce J, Ferrier S (2000) Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling 133(3):225-245

Pearson RG, Phillips SJ, Loranty MM et al (2013) Shifts in Arctic vegetation and associated feedbacks under climate change. Nature Climate Change

Peterson AT, Holt RD (2003) Niche differentiation in Mexican birds: using point occurrences to detect ecological innovation. Ecology Letters 6(8):774-782

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. Ecological Modelling 190(3):231-259

Phillips SJ, Dudík M, Elith J et al (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol. Appl. 19(1):181-197

Phillips SJ, Elith J (2010) POC plots: calibrating species distribution models with presence-only data. Ecology 91(8):2476-2484

Phillips SJ, Elith J (2013) On estimating probability of presence from use-availability or presence-background data. Ecology 94(6):1409-1419

Reese GC, Wilson KR, Hoeting JA, Flather CH (2005) Factors affecting species distribution predictions: a simulation modeling experiment. Ecol. Appl. 15(2):554-564

Ricklefs RE, Jenkins DG (2011) Biogeography and ecology: towards the integration of two disciplines. Philosophical Transactions of the Royal Society B: Biological Sciences 366(1576):2438-2448

Royle JA, Chandler RB, Yackulic C, Nichols JD (2012) Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. Methods in Ecology and Evolution 3(3):545-554

Sax DF, Early R, Bellemare J (2013) Niche syndromes, species extinction risks, and management under climate change. Trends in ecology & evolution 28(9):517-523

Shen Z, Fei S, Feng J et al (2012) Geographical patterns of community-based tree species richness in Chinese mountain forests: the effects of contemporary climate and regional history. Ecography 35(12):1134-1146

Václavík T, Meentemeyer RK (2012) Equilibrium or not? Modelling potential distribution of invasive species in different stages of invasion. Diversity and Distributions 18(1):73-83

Vaughan I, Ormerod S (2005) The continuing challenges of testing species distribution models. Journal of Applied Ecology 42(4):720-730

Wisz MS, Hijmans R, Li J, Peterson AT, Graham C, Guisan A (2008) Effects of sample size on the performance of species distribution models. Divers. Distrib. 14(5):763-773

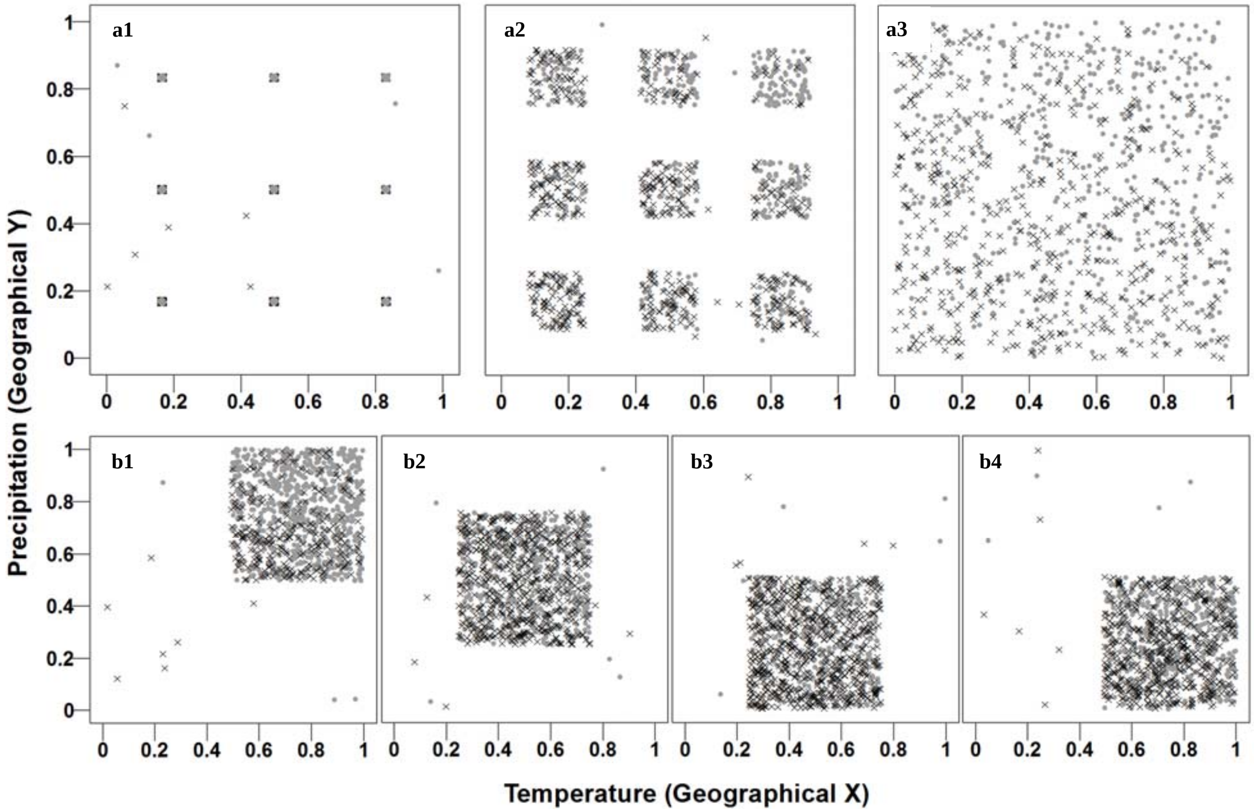


Figure 1. Distribution of sampling points (dots represent species presence and crosses represent pseudo-absence) of a virtual species in an environmental space. (a) Examples of point pattern gradient from highly clustered (a1 - Dataset 1) and medium clustered (a2 - Dataset 25) to randomly distributed (a3 - Dataset 50) in data series A; (b) examples of point pattern locations of medium clustered data (Dataset 25) in Series B (b1), Series C (b2), Series D (b3), and Series E (b4). See Appendix 1 for detailed spatial arrangement and methodology.

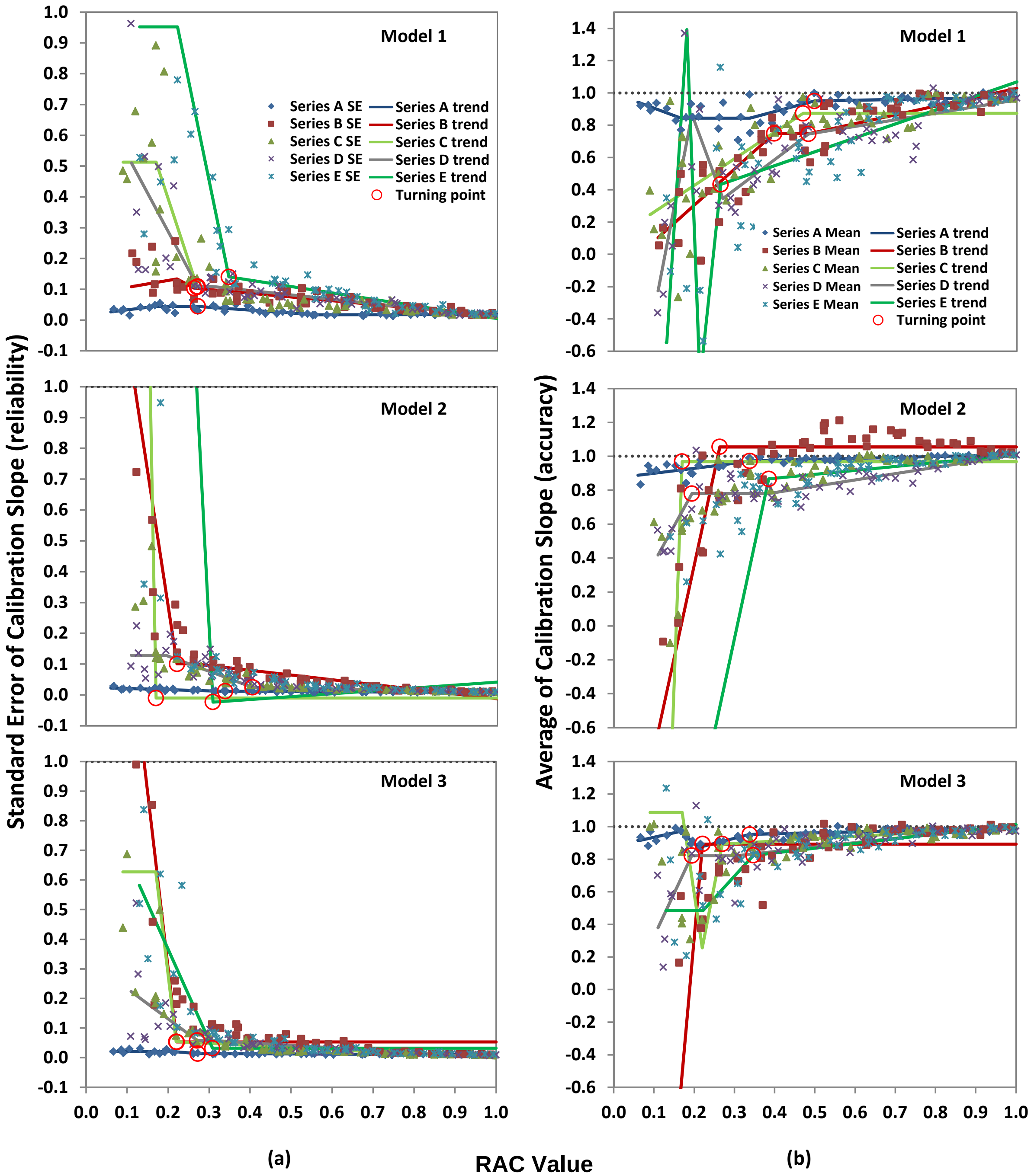


Figure 2. The relationships between RAC value and (a) model reliability and (b) model accuracy by data series and model. Model reliability is measured by standard error of the regression slope of the binned method (see Appendix 3 for details), while model accuracy is measure by the mean of the regression slope.

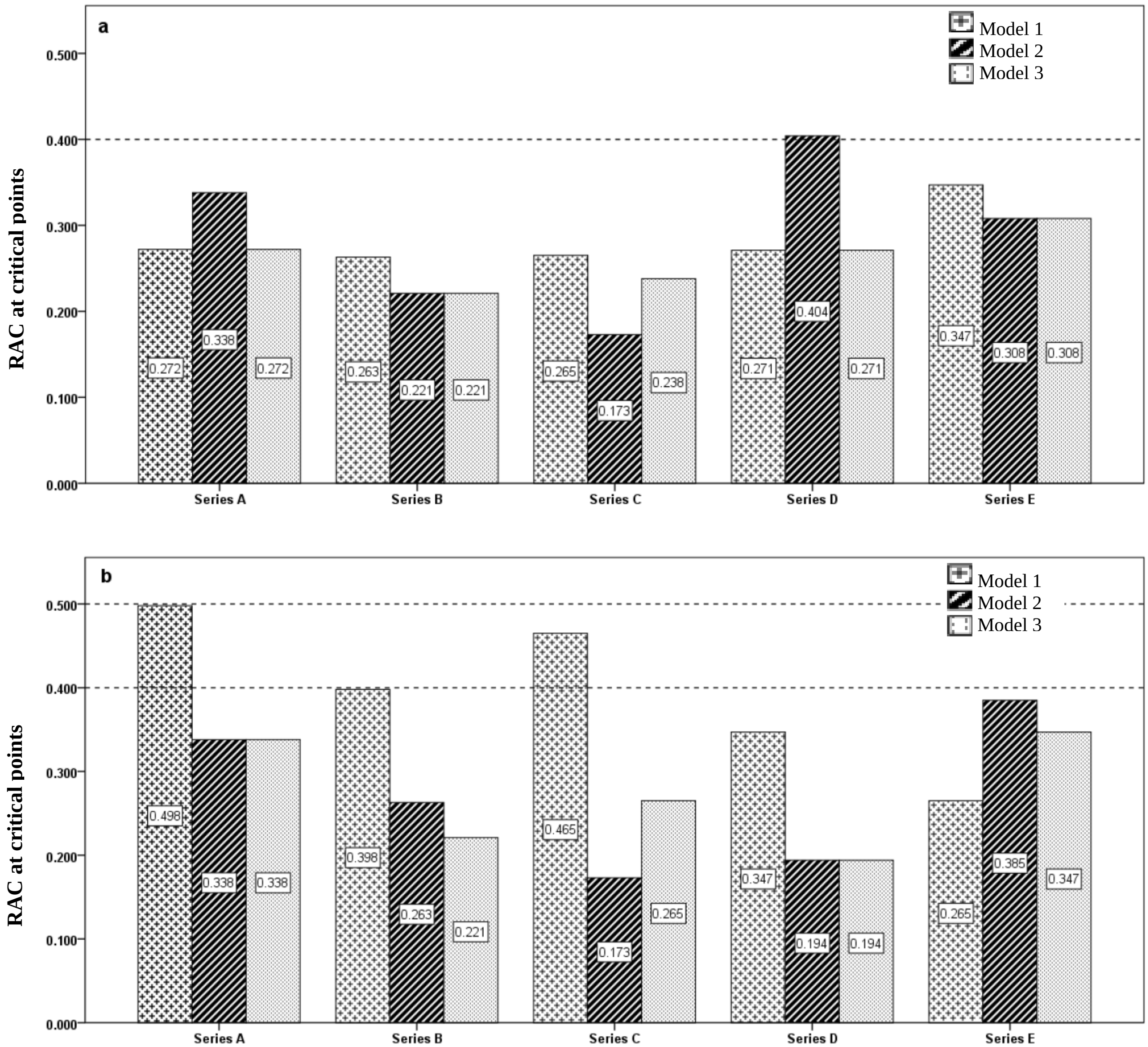


Figure 3. RAC value at the critical points when model reliability (a) and model accuracy (b) are stabilized for the three models in each of the five point pattern series in the environmental space.

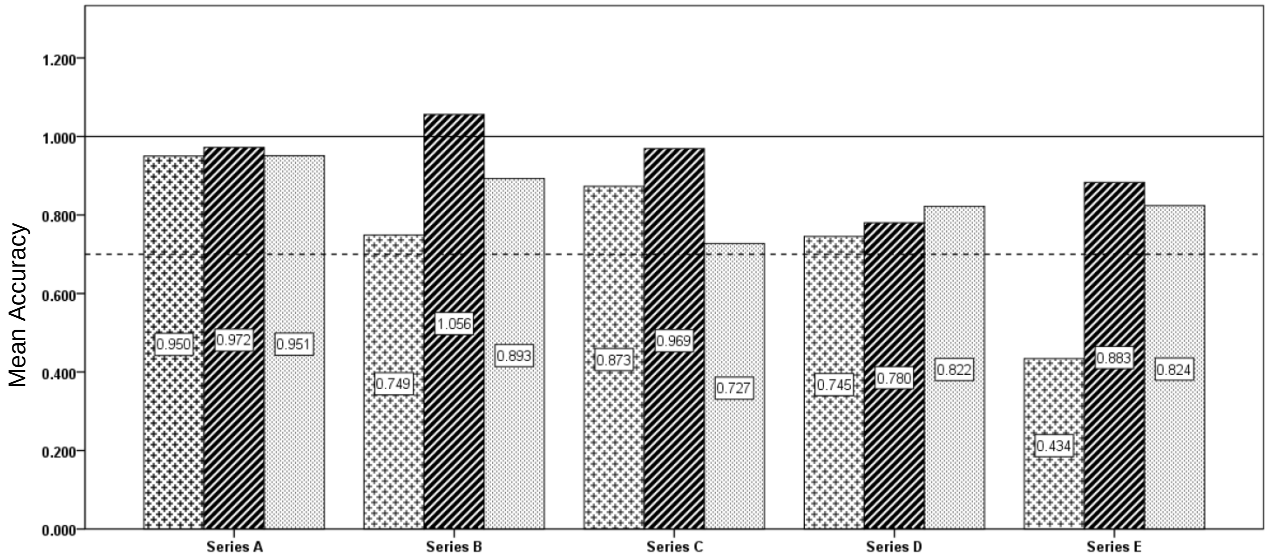


Figure 4. Mean accuracy, as measured by the slope of the regression line in the binned calibration method, at the critical threshold by model type and data series.

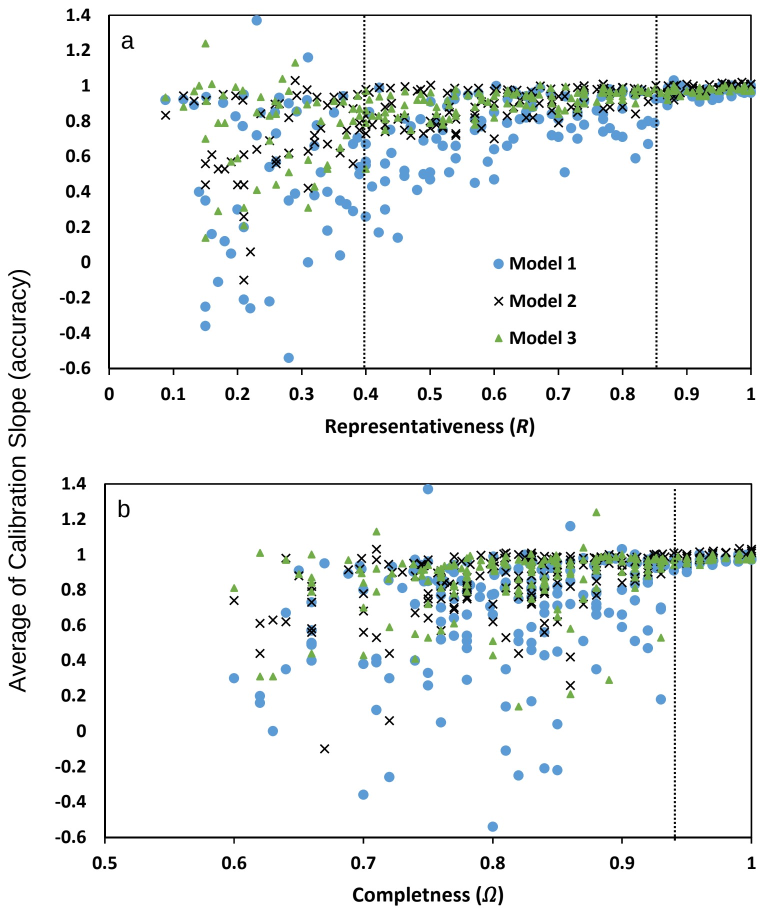
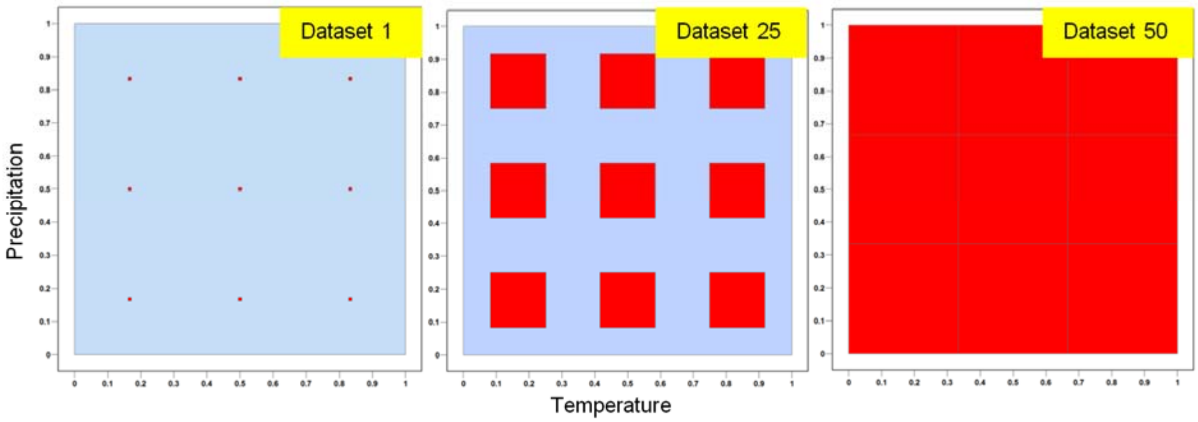


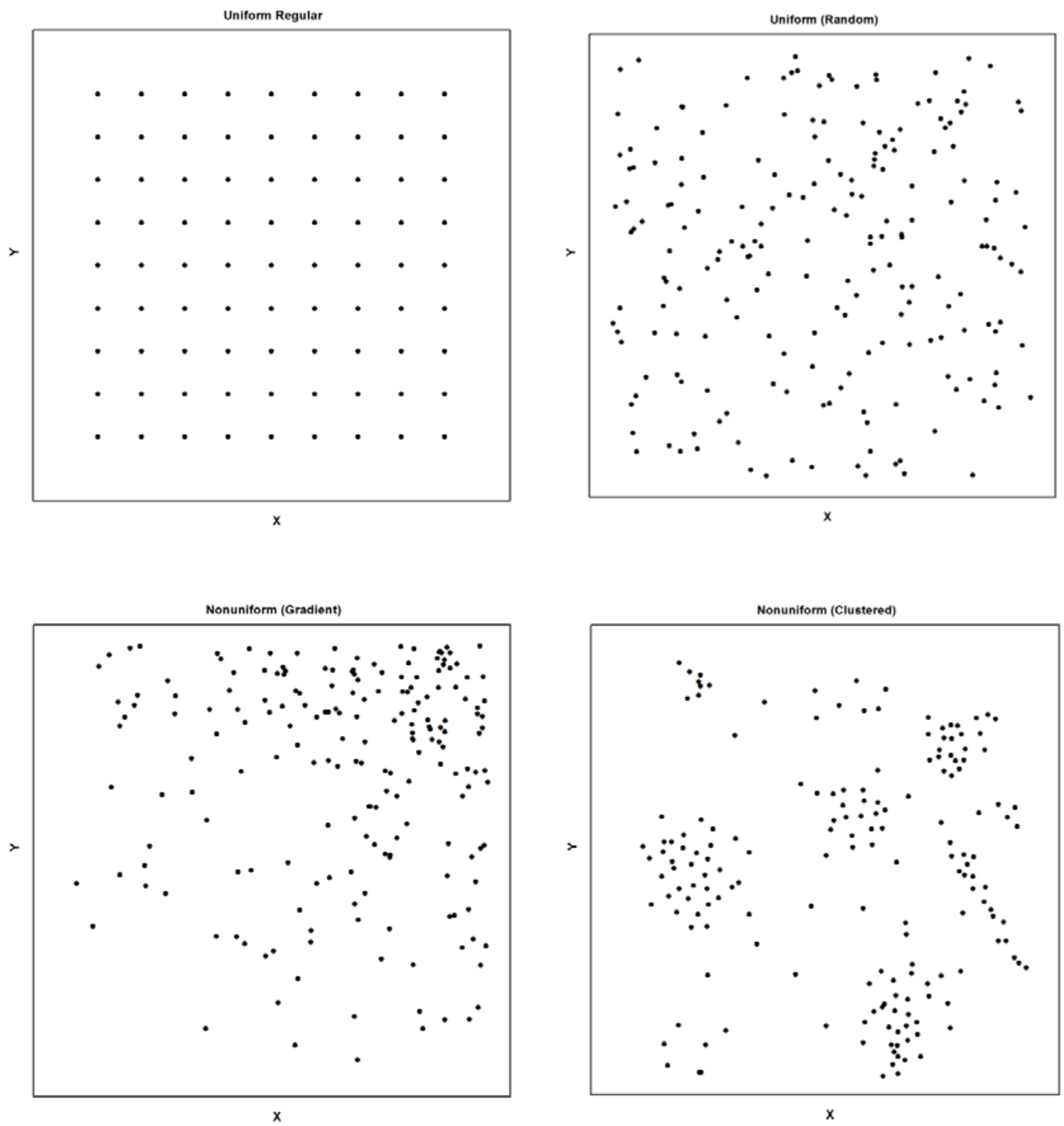
Figure 5. The relationships between model accuracy (as measured by the slope of the regression line in the binned calibration method) and (a) data representativeness (*R*) and (b) data completeness (*Ω*) by prediction model.

Appendix 1. Summary of distribution center, spatial arrangement, and location of the virtual species in the environmental space of the five point pattern series A – E.
In each series, dataset 1 is highly clustered (non-uniform) and dataset 50 is completely random. In series A, Z1 is the combination of the nine grid-squares evenly distributed on the environmental space. The size of each grid unit gradually increases without shifting their centers until they covered the entire environmental space at dataset 50. Meanwhile, Z2 (the complimentary subareas of Z1 (total area – Z1)) simultaneously decreases in size until it vanishes at dataset 50. In series B to E, Z1 is a random distribution in one square unit with the size equivalent to the total size of the nine grid-square units in series A. Z1 evolves from the respective center coordinates specified in the following table and then gradually increases in size until it covers the entire environmental space. Z2 simultaneously decreases in size until vanishes at dataset 50.

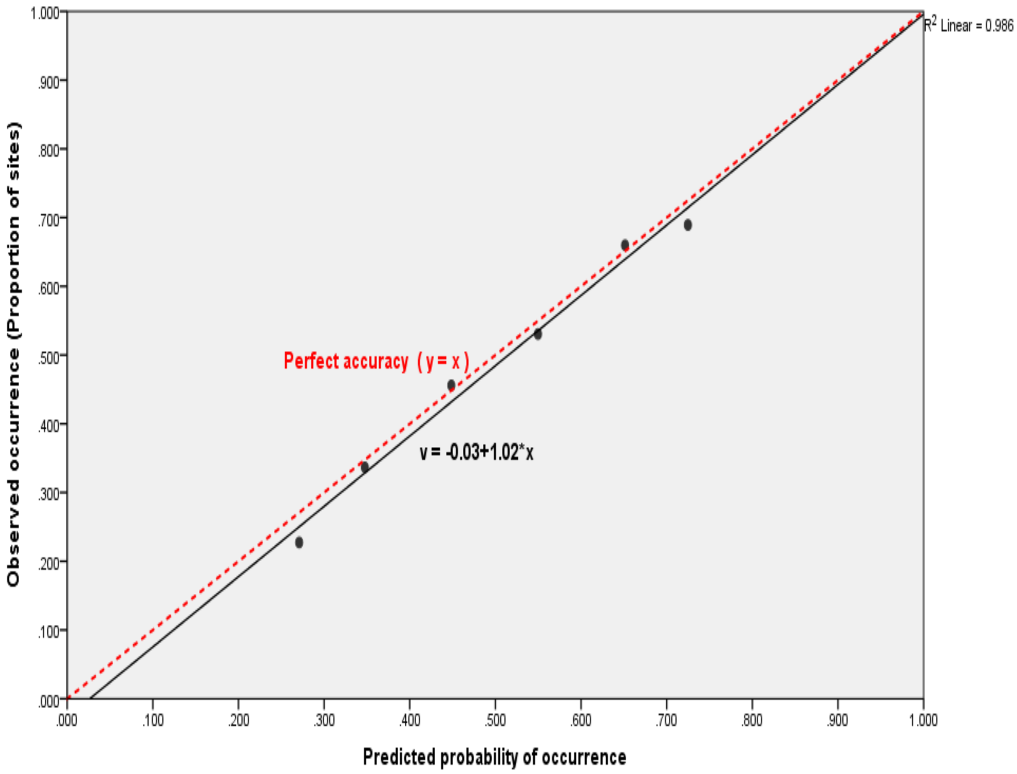


Series	Distribution Center of Z1	Description of spatial arrangement and location of Z1 in relation to the edges of the environmental space
A	Nine grids centers, evenly distributed in the landscape (0.167, 0.167), (0.167, 0.500), (0.167, 0.833). (0.500, 0.167), (0.500, 0.500), (0.500, 0.833), (0.833, 0.167), (0.833, 0.500), (0.833, 0.833)	Evolved from Multi-center clustered to random pattern Z1: Not touching the edges
B	Top right corner of the landscape (0.985, 0.985)	Evolved from Mono-center clustered to random pattern Z1: touching edges for both variables, both with top value
C	Center of the landscape (0.500, 0.500)	Evolved from Mono-center clustered to random pattern Z1: Not touching the edges
D	Shift from bottom to center of the landscape (0.500, 0.150) to (0.500, 0.500)	Evolved from Mono-center clustered to random pattern Z1: Toughing the edge of the Temperature variable
E	Bottom right corner of the landscape (0.985, 0.015)	Evolved from Mono-center clustered to random pattern Z1: Touching the edges for both variables, one with top value (Temperature)

Appendix 2. Visualization of basic types of point patterns. Uniform pattern describes the spatial pattern in which the density of points in any subarea of the interested region is equal if the size and shape being the same. Conversely, non-uniform pattern describes the pattern with the density of points varies between one subarea and the other. Quantitatively, a regular point pattern (a) means the distance between each point and their paired points remains the same for one or more specified directions (in this case vertical and horizontal) within the focused area; while a random point pattern (b) means the probability of containing a point in one subarea is the same as any other subarea of the same size and shape, regardless of the location of these subareas. A gradient pattern means (c) the probability of locating one individual point varies inversely with distance to the points have already been located from single-clustered center; while a clustered pattern (d) means probability of locating one individual point varies inversely with distance to the points have already been located, but allowed to have multi-clustered centers.



Appendix 3 . Binned calibration method. Red line corresponds to the perfect calibration of the model, when the plotted points fall on the 1:1 line. The coefficient of the regression line (black) represents the overall calibration of each run from the individual datasets (1 to 50) of each series



Appendix 4 Point patterns of species occurrence data in the extracted environmental space (from the native range) for *Prosopis farcta* (A) and *Imperata cylindrica* (B), represented by the component scores of the dimension reduction results of PCA. Red dots represent the occurrence data locations reflected in the environmental space. Boundaries of the environmental space correspond to their environmental space within the native ranges. The extents of the environmental space reflect the relative size of the native ranges for these two species