ISSN 0104-7760
v.26 n.2 2020

Cheng-Tao Lin[1a], Ching-An Chiu[2a+]

COMPARISON OF PREDICTOR SELECTION PROCEDURES IN SPECIES DISTRIBUTION MODELING: A CASE STUDY OF *Fagus hayatae*

LIN, C. T.; CHIU, C. A. Comparison of predictor selection procedures in species distribution modeling: a case study of *Fagus hayatae*. **CERNE**, v. 26, n. 2, p.172-182, 2020.

HIGHLIGHTS

Too many or too few environmental variables are unsuitable for SDM.

Ineffective indicators for species distribution must be excluded

Bioclimatic indicators are key plant distribution variables.

Collinearity, contribution, and ecological significance affect predictor selection

This synthetic strategy for integrating correlation coefficient, contribution level, and expert choice of predictors can facilitate the selection of suitable environmental factors..

Keywords:
Synthetic strtegy
Effective warmth index
Expert knowledge
Environmental variableTaiwan

ABSTRACT

Selecting predictors for species distribution models (SDMs) is a major challenge. In this study, we evaluated a comprehensive set of 62 environmental predictors that may be related to the occurrence of Fagus hayatae. We modeled F. hayatae as a case study to compare model performance through different environmental predictor subsets according to three selection procedures, namely correlation coefficients between predictors, contribution level of predictors, and expert choice of biologically relevant predictors. The three selection procedures provided satisfactory results with high performance using about 6-10 valid predictors but had their respective limitations. Consequently, we suggest a synthetic strtegy of predictor selection. Accordingly, the first step was identifying and eliminating ineffective variables with nonidentifiability by using bivariate scatterplots. Next, calculate the correlation coefficients between other candidate predictors. Finally, comprehensively select the applicable environmental predictors with lower correlation coefficient on the basis of highly contribution level and expert knowledge for SDM of target species.

Historic:
Received 28/07/2019
Accepted 13/05/2020

+Correspondence:
cachiu@nchu.edu.tw

DOI:
10.1590/01047760202026022657

[1] National Chiayi University, Chiayi, Taiwan - ORCID:0000-0003-2857-1625[a]
[2] National Chung Hsing University, Taichung, Taiwan - ORCID: 0000-0002-6134-0530[a]

INTRODUCTION

In the past three decades, species distribution models (SDMs) have been increasingly used to solve many scientific and managerial problems related to climate change, conservation planning, ecological theory, and invasive species (Guisan and Thuiller, 2005; Austin and Van Niel, 2011; Petitpierre et al., 2017; Zhang et al., 2019). Generally, SDMs are composed of three elements (Sangermano and Eastman, 2012), namely a dependent variable (species occurrence data), explanatory variables (environmental predictors), and an algorithm or function for representing species–environment relationships (modeling methods). Numerous studies have explored the performance of different modeling methods (Guisan et al., 2007) and effects of species sampling sizes and spatial bias (Wisz et al., 2008; Syfert et al., 2013). By contrast, environmental predictors have been less commonly discussed (Franklin, 2010), although the selection of environmental predictors is relevant to SDM performance and its subsequent applications (Williams et al., 2012). Thus, the appropriate selection of predictors for SDMs is a major challenge (Araújo and Guisan, 2006; Franklin, 2010; Watling et al., 2012; Petitpierre et al., 2017).

To the greatest possible extent, data of species occurrence and environmental predictor layers should be collected before modeling species distribution. Environmental predictors can be divided into indirect, direct, and resource gradients (Guisan and Thuiller, 2005). According to the conceptual model of using environmental predictors (Guisan and Zimmermann, 2000; Franklin, 2010), direct and resource variables are preferred for predicting plant distribution. For example, Austin and Van Niel (2011) suggested that light, temperature, nutrients, water, CO_2 levels, disturbance, and biota are the seven groups of variables that control plant distribution. Although attention should be mainly focused on explanatory power and the ecological basis of choosing predictors (Araújo and Guisan, 2006), the use of candidate variables of an SDM depends on the availability of environmental layers. In practice, many environmental predictors used in SDMs are indirect variables or surrogates of direct and resource variables (Franklin, 2010; Austin and Van Niel, 2011); for example, elevation is used as an agency of mountain temperature based on lapse rate (Chiu et al., 2014).

Problems related to over-parameterization and overfitting of a model may arise with the use of and excessive number of variables in SDMs (Guisan and Zimmermann, 2000; Tyberghein et al., 2012), particularly when highly correlated variables exist. Collinearity or high correlation between variables is a common feature of ecological data (Mac Nally, 2002). Adopting a predictor selection procedure to reduce the number of variables to as few as possible is necessary for two reasons (Dormann, 2011). First, a higher number of variables corresponds to greater correlation among them. Second, a higher number of variables corresponds to a greater likelihood of one of them spuriously contributing to the model (type I error). Although principal component analysis is a method used to deal with collinearity of variables (Dormann et al., 2013), it is difficult to interpret a modeled result based on it. Alternatively, some studies have reported retaining only those predictors that contribute more than 0.5% or 5% to the SDM and discarding the others (Young, 2010; Yang et al., 2013). However, to avoid the collinearity or nonidentifiability of predictors, a common method used is identifying and eliminating redundant variables by using the pairwise Pearson correlation coefficients among all candidate variables. The absolute value of the threshold of the correlation coefficient ($|r|$) is usually 0.7 (Dormann et al., 2013); however, different researchers may use different threshold values, such as $|r| > 0.9$ (Randin et al., 2009), $|r| > 0.85$ (Syfert et al., 2013), $|r| > 0.8$ (Young, 2010), and $|r| > 0.6$ (Andreo et al., 2011). In addition to noncorrelated predictors, Watling et al. (2012) used a subset of user-defined biologically relevant environmental predictors in SDMs. Consequently, Williams et al. (2012) suggested that the identification of redundant variables relies on a combination of *a priori* ecological considerations, knowledge of derivation and accuracy of each variable, awareness of relationships among variables, and a rigorous process of testing the utility of alternative sets of predictors in a statistical model.

The appropriate selection of environmental predictors is a critical step in modeling species distribution (Araújo and Guisan, 2006; Austin and Van Niel, 2011; Watling et al., 2012; Williams et al., 2012). Three common approaches of variable selection are as follows: using all available bioclimatic variables without justification, reducing the number of bioclimatic and biophysical covariates to account for collinearity, and selecting variables on the basis of ecological knowledge (Porfirio et al., 2014). Each SDM has an appropriate predictor selection method associated with it, namely expert knowledge, statistical significance, and iterative selection based on training accuracy (Lippitt et al., 2008). In this study, we compared the performance of models by using different environmental predictor subsets corresponding to three selection procedures, namely correlation coefficients between predictors, contribution levels of predictors, and expert selection of biologically relevant predictors.

MATERIAL AND METHODS

Study area and target species occurrence data

Taiwan is a subtropical mountainous island at the periphery of East Asia (Figure 1A). The study area covered an area in northern Taiwan (Figure 1B), extending from 121.1925–121.8913°E and 24.4210–24.9024°N. The target species in this modeling study was *Fagus hayatae* Palib., a relic and stenotopic tree. The distribution area of *F. hayatae* was mapped in the National Vegetation Diversity Inventory and Mapping Project of Taiwan (Chiou et al., 2009), extending over 1,282 ha and ranging from 1,100 to 2,100 m above sea level and limited to a few ridges nearby (Figure 1B). We generated regular points at 200-m intervals by using ArcGIS version 10.0 (ESRI; Redlands, USA) with the Geospatial Modeling Environment software version 0.7.2.0 (Beyer, 2012). According to our species occurrence data, 319 *F. hayatae* points were extracted (presence-only; Figure 1C) and used in all of the SDMs with different environmental predictor subsets. The regular 319 occurrence data prevented the effects of sample size and spatial bias for SDMs (Wisz et al., 2008; Syfert et al., 2013).

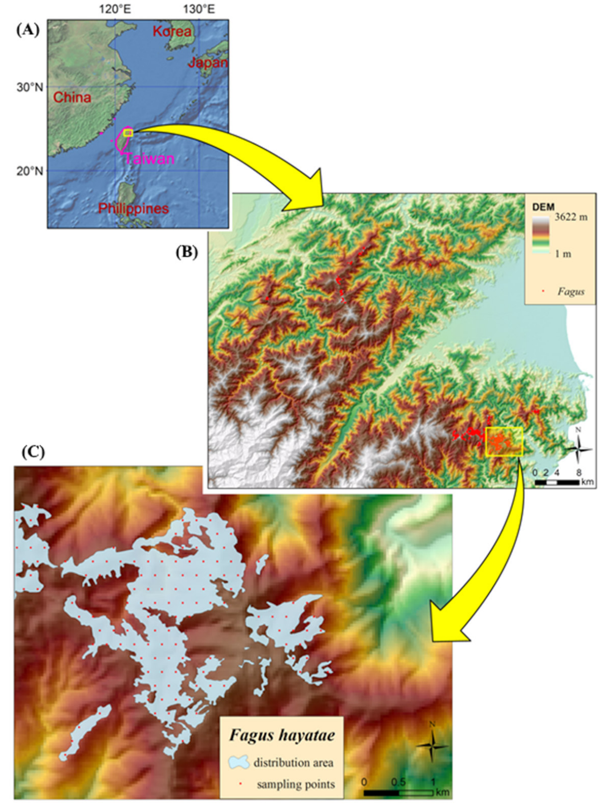

FIGURE 1 (A) Geographic location of Taiwan. (B) Digital elevation model of the study area and spatial distribution (red points) of *Fagus hayatae*. (C) Partially enlarged occurrence data (red sampling points) of *F. hayatae*.

Environmental predictors

We evaluated a comprehensive set of 62 environmental predictors that may be related to the occurrence of *F. hayatae*. The 62 predictors and their abbreviations are listed in Supplementary Table 1. These include monthly mean temperature and precipitation (Chiu et al., 2009), bioclimatic variables (version 1.4; Hijmans et al., 2005), warmth index, coldness index (Kira, 1991), biotemperature, potential evapotranspiration ratio (Holdridge, 1967), summer and winter half-yearly precipitation (Su, 1985), humidity index (Xu, 1985), effective warmth index (Chiu et al., 2012), temperature annual range (Wolfe, 1979), whole light sky space (Lai et al., 2010), elevation, latitude, longitude, and slope, which were produced using a digital elevation model (DEM) provided by the Taiwan Forestry Bureau Aerial Survey Office. Furthermore, by using the Geomorphometry and Gradient Metrics Toolbox (Evans, 2011), data concerning dissection, roughness, compound topographic index, heat load index, topographic radiation aspect index, surface relief ratio, surface curvature index, slope position, and surface/area ratio were calculated from DEM. All environmental predictor layers were generated in ArcGIS at a 40-m spatial resolution in the same geographic extent.

Selection procedure for environmental predictors

Before selecting environmental predictors, we eliminated the variables without discrimination power, which were identified through detection using bivariate scatterplots in the preliminary model analysis and through decision-making (Morisette et al., 2013).

In this study, we used the following three procedures with backward elimination to select environmental predictors on the basis of comprehensive recommendations made in different studies, including studies by Lippitt et al. (2008), Dormann (2011), Watling et al. (2012), Williams et al. (2012), and Porfirio et al. (2014).

1. Selection based on the Pearson correlation coefficients (r) between predictors: Highly correlated variables (following the sequence $|r| > 0.95$, 0.90, 0.85, 0.80, 0.75, 0.70, 0.65, 0.60, and 0.55) were removed, and others were retained as environmental predictors in the SDM. In this procedure, if 61 predictors were retained and used in the SDM, we labeled the predictor subset as R61, and the subset with 32 predictors was labeled as R32; a similar labeling method was adopted for all subsets.

2. Selection based on the training contribution level of predictors: Variables with low contribution percentages (CP) of <0.1%, 0.2%, 0.3%, 1%, 3%,

and 9% were sequentially removed, and others were retained as environmental predictors in the SDM. In this procedure, if 61 predictors were retained and used in the SDM, we labeled the predictor subsets as C61, and the subset with 30 predictors was labeled as C30; a similar labeling method was employed for all subsets.

3. Selection based on expert selection of biologically relevant predictors: Five experienced ecologists studied *F. hayatae* and removed the biologically irrelevant variables; the other variables were retained as environmental predictors in the SDM. In this procedure, if 61 predictors were retained and used in the SDM, we labeled the predictor subsets as E61, and the subset with 31 predictors was labeled as E31; a similar labeling method was adopted for all subsets.

Model creation and evaluation of accuracy

To predict the distribution of *F. hayatae*, we used a general-purpose machine learning SDM method, Maximum Entropy Modeling of Species Geographic Distributions (MaxEnt; Elith et al., 2011). MaxEnt has been shown to be a robust SDM method for presence-only species data (Guisan et al., 2007), and it is becoming one of the most widely used methods for SDMs. We used MaxEnt 3.3.3 k (http://www.cs.princeton.edu/~schapire/maxent/) for 20 cross-validated replicates, 5000 maximum iterations, and logistic output format with other default settings. The logistic output was used as a species-suitable index value or predicted occurrence probability (Elith et al., 2011).

To assess the performance of models using different environmental predictor subsets, we used receiver operating characteristic (ROC) curve analysis and the true skill statistic (TSS) (Allouche et al., 2006). Use of the ROC is a common approach for threshold independent evaluation, where "sensitivity" is plotted against "1 − specificity" for all possible thresholds, thus avoiding the subjective selection of one or several thresholds for the evaluation (Gontier et al., 2010). Following the guidelines from Swets (1988), standard values for the area under the curve (AUC) of the ROC plots were graded as follows for assessing model performance: failed (AUC = 0.5–0.6), poor (AUC = 0.6–0.7), fair (AUC = 0.7–0.8), good (AUC = 0.8–0.9), and excellent (AUC = 0.9–1.0). In addition, TSS is a threshold-dependent evaluation index. We adapted the suggestion provided in a study by Liu et al. (2013) and used the maximized sum of sensitivity and specificity as the threshold to transform the predictive occurrence probability into species presence or absence (1 or 0). For a 2 × 2 confusion matrix, TSS was defined (Allouche et al., 2006) as $TSS = (ad - bc)/[(a + c)(b + d)] =$ sensitivity + specificity – 1.

RESULTS AND DISCUSSION

Bivariate scatterplot between all predictors

A bivariate scatterplot is a preliminary inspection that reveals the relationship between paired environmental predictors. Figure 2A shows all scatterplots created in ArcGIS and some examples of the paired predictors. The random relationship between TAR (for abbreviations, refer to Supplementary Table 1) and DIS predictors, where the correlation coefficient was <0.001, is presented in Figure 2B. High correlation, or collinearity, was observed between ELE and Bio1 (Figure 2C); the correlation coefficient was 0.998. If a model contains collinear predictors, the variance estimates are typically inflated, resulting in biased prediction (Quinn and Keough, 2002; Solomon et al., 2002). The scatterplot of CI vs. TRAI (Figure 2D) clearly showed no discrimination of CI variable with constant zero values. Although CI variables have often been used to indicate the upper limit of the distribution of evergreen broad-leaved forests, their use is inappropriate for a subtropical island such as Taiwan (Chiu et al., 2014). After examining each scatterplot of all paired environmental predictors, we discarded the CI variable due to its nonidentifiability and retained the other 61 variables as candidate predictors in subsequent processing procedures. Figure 2 also reveals some collinearity among environmental predictors (see Supplementary Table 2). Consequently, creating the scatterplot for each pair of environmental predictors before running the SDM was correct and necessary (Williams et al., 2012; Morisette et al., 2013).

Predictor subsets through backward elimination

Table 1 presents the subsets of selected predictors used in different SDMs through backward elimination. These predictor subsets were labeled using three selection procedures; subsets were labeled with "R" for Pearson correlation coefficient (r), "C" for training contribution level, and "E" for expert selection, followed by the number of predictors used in different subsets. Supplementary Table 2 lists all pairwise Pearson correlation coefficients between predictors. Ten predictor subsets were based on Pearson correlation coefficients of predictors used in SDMs; these included R61, R32, R27, R23, R21, R18, R12, R9, R8, and R6. Furthermore, seven predictor subsets based on the training contribution level of predictors used in SDM were available, namely C61, C30, C28, C23, C15, C9, and C6. In addition, seven predictor subsets based on expert selection of biologically relevant predictors were used in SDM, namely E61, E31, E26, E21, E16, E11,

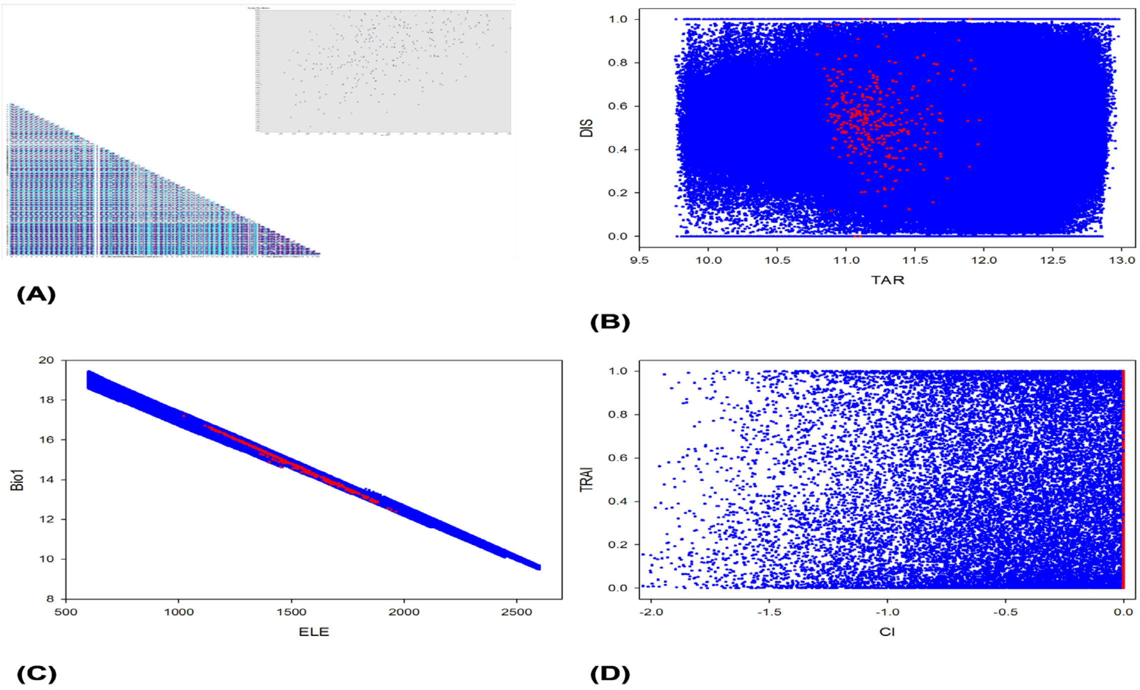

FIGURE 2 Examples of scatterplots of variables: (A) scatterplot matrix between all environmental predictors generated by AcrGIS; (B) TAR vs. DIS plot (for abbreviations, refer to Supplementary Table 1), which reveals a random relationship (|r| < 0.001); (C) ELE vs. Bio1 plot, which presents high correlation (|r| > 0.998); (D) CI vs. TRAI plot, which indicates no discrimination between CI values in sampling points. Blue points represent the values of environmental predictors on all modeling grids; red points represent the values of environmental predictor on all Fagus hayatae occurrence grids.

and E6. The ratio of the maximum (R61, C61, and E61) to minimum (R6, C6, and E6) number of predictors in different subsets was > 10.

Table 2 presents the r between predictors and the CP (%) of individual predictors used in the R6, C6, and E6 models. Only PER and CTI were selected twice, whereas the other predictors were selected only once in the R6, C6, and E6 models. The results revealed that different selection procedures resulted in different predictor subsets used in SDM. Italicization indicates high correlation between T6 and Bio5 (|r| = 0.90) and between Bio16 and longitude of raster (LON) (|r| = 0.93) in the C6 model and between Bio4 and WI (|r| = 0.88) in the E6 model. All predictors in the R6, C6, and E6 models could be divided into following two categories: thermal/moisture-related or topography/geography-related variables. The CP of individual predictors from 46.7 to 0.0 exhibited a high level of inconsistency among models such as R6, C6, and E6 (Table 2).

Performance comparison of models on the basis of different predictor subsets

Figure 3 presents a comparison of model performance measured using AUC and TSS criteria within 10-R-subsets, 7-C-subsets, and 7-E-subsets. The AUC was excellent (> 0.9) across all 22 models with different predictor subsets (Supplementary Table 3).

Overall, the AUC values (overall mean 0.976) were higher than the TSS (overall mean: 0.939), as shown in Figure 3. Analysis of variance indicated significant differences (P < 0.05) within the 10-R-subsets, 7-C-subsets, and 7-E-subsets for the ROC and TSS criteria. Only minor differences were observed among the models that used 12 or more predictors with respect to both AUC and TSS, namely R61–R12 (Figure 3A and B), C61–C15 (Figure 3C and D), and E61–E16 (Figure 3E and F). The five models, namely R9, R8, R6, E11, and E6, used as predictor subsets (Figure 3A, B, E and F) had significantly lower performance than the other subsets. Figure 3 also presents a significant reduction in SDM accuracy between R12 and R9, C15 and C9, and E11 and E6. The results suggest that in our case, the appropriate number of predictors was approximately 10. Moreover, satisfactory performance could be achieved using as few as six predictors (Figure 3C and D). In a review by Porfirio et al. (2014), 119 variables were once used in different SDM studies. The mean annual precipitation and mean annual temperature were the most commonly used variables, observed in 43% and 37% of related studies, respectively. In this study, variables related to temperature and moisture or their integrated index were used in the R6, C6, and E6 models, which again indicated the crucial roles of temperature and water in plant distribution (Austin and Van Niel, 2011).

TABLE 1 The total 22 subsets of predictors selected through backward elimination used in different SDM models, labeled with three selection procedures (Pearson correlation coefficient as "R," training contribution level as "C," and expert selection as "E") and appended with the number of predictors.

Label	Subsets of selected predictors
R61 = C61 = E61	T1, T2, T3, T4, T5, T6, T7, T8, T9, T10, T11, T12, P1, P2, P3, P4, P5, P6, P7, P8, P9, P10, P11, P12, Bio1, Bio4, Bio5, Bio6, Bio8, Bio9, Bio10, Bio12, Bio15, Bio16, Bio17, Bio18, BT, PER, WI, EWI, HI, PS, PW, PSR, PWR, TAR, ELE, SLO, LAT, LON, SR, DIS, ROU, SRR, CUR, SP, SAR, CTI, HLI, TRAI, WLS
R32	P2, P3, P4, P5, P6, P7, P8, Bio4, Bio5, Bio6, Bio12, Bio15, Bio16, Bio17, Bio18, PER, WI, HI, PS, PSR, SLO, LAT, LON, SR, DIS, SRR, CUR, SAR, CTI, HLI, TRAI, WLS
R27	P2, P3, P5, P6, P7, P8, Bio4, Bio5, Bio6, Bio15, Bio16, Bio17, Bio18, PER, WI, HI, PS, LAT, SR, DIS, SRR, CUR, SAR, CTI, HLI, TRAI, WLS
R23	P2, P3, P5, P6, P8, Bio4, Bio6, Bio15, Bio16, Bio17, Bio18, PER, PS, LAT, SR, DIS, SRR, CUR, SAR, CTI, HLI, TRAI, WLS
R21	P3, P5, P6, P8, Bio4, Bio6, Bio15, Bio16, Bio17, Bio18, PER, PS, LAT, SR, SRR, CUR, SAR, CTI, HLI, TRAI, WLS
R18	P5, Bio4, Bio6, Bio15, Bio16, Bio17, Bio18, PER, PS, LAT, SR, SRR, CUR, SAR, CTI, HLI, TRAI, WLS
R12	Bio4, Bio15, Bio16, Bio18, PER, SR, SRR, CUR, SAR, CTI, TRAI, WLS
R9	Bio4, Bio15, Bio18, PER, SR, SRR, SAR, CTI, WLS
R8	Bio4, Bio15, Bio18, PER, SR, SRR, SAR, CTI
R6	Bio15, Bio18, PER, SR, SRR, CTI
C30	T1, T6, T7, T8, P1, P2, P4, P5, P6, P8, P9, P10, P11, Bio4, Bio5, Bio6, Bio9, Bio10, Bio15, Bio16, Bio17, Bio18, HI, PS, PW, TAR, ELE, LAT, LON, TRAI
C28	T6, T7, T8, P1, P2, P4, P5, P6, P8, P9, P10, P11, Bio4, Bio5, Bio6, Bio9, Bio10, Bio15, Bio16, Bio17, Bio18, HI, PS, TAR, ELE, LAT, LON, TRAI
C23	T6, T7, T8, P1, P2, P8, P9, P10, P11, Bio4, Bio5, Bio10, Bio15, Bio16, Bio17, Bio18, HI, PS, TAR, ELE, LAT, LON, TRAI
C15	T6, T7, T8, P1, P8, Bio4, Bio5, Bio15, Bio16, Bio18, HI, PS, LAT, LON, TRAI
C9	T6, T7, P1, Bio5, Bio15, Bio16, HI, PS, LON
C6	T6, P1, Bio5, Bio16, PS, LON
E31	Bio4, Bio5, Bio6, Bio8, Bio9, Bio10, Bio12, Bio15, Bio16, Bio17, Bio18, PER, WI, EWI, HI, PS, PW, PSR, PWR, SLO, SR, DIS, ROU, SRR, CUR, SP, SAR, CTI, HLI, TRAI, WLS
E26	Bio4, Bio5, Bio6, Bio8, Bio9, Bio10, Bio12, Bio15, Bio16, Bio17, PER, WI, EWI, HI, PS, PW, SLO, SR, DIS, ROU, CUR, SP, SAR, CTI, TRAI, WLS
E21	Bio4, Bio5, Bio6, Bio8, Bio9, Bio12, Bio15, Bio17, PER, WI, EWI, PS, PW, SLO, SR, CUR, SP, SAR, CTI, TRAI, WLS
E16	Bio4, Bio5, Bio8, Bio12, Bio15, Bio17, PER, WI, EWI, PS, SLO, SR, SP, CTI, TRAI, WLS
E11	Bio4, Bio8, Bio12, Bio17, PER, WI, EWI, SR, SP, CTI, WLS
E6	Bio4, Bio12, PER, WI, CTI, WLS

TABLE 2 Pearson correlation coefficient (r) between predictors and their contribution percent (CP, %) individually in the R6, C6, and E6 models. Italicization indicates a strong correlation between predictors.).

R6 model	Bio15	Bio18	PER	SR	SRR	CTI
Bio15		-0.34	-0.09	-0.02	0.00	-0.01
Bio18			0.23	0.00	0.00	0.02
PER				-0.22	-0.08	0.07
SR					0.08	0.08
SRR						-0.05
CTI						
CP (%)	33.8	18.2	46.7	1.1	0.1	0.0
C6 model	T6	P1	Bio5	Bio16	PS	LON
T6		0.40	0.90	0.32	0.36	0.37
P1			0.34	0.59	0.53	0.56
Bio5				0.30	0.34	0.36
Bio16					0.68	0.93
PS						0.59
LON						
CP (%)	22.7	12.7	10.6	28.4	11.6	14.0
E6 model	Bio4	Bio12	PER	WI	CTI	WLS
Bio4		0.20	0.54	0.88	0.08	0.01
Bio12			-0.62	0.42	0.00	0.17
PER				0.44	0.07	-0.20
WI					0.09	-0.03
CTI						-0.16
WLS						
CP (%)	14.9	38.2	20.2	26.3	0.1	0.4

As indicated in Figures 3A and B, setting the correlation coefficient threshold between variables at 0.7 was appropriate (|r| > 0.7 variables eliminated in R12 with higher accuracy, |r| > 0.65 variables eliminated in R9 with lower accuracy). The |r| = 0.7 threshold was recommended by Dormann et al. (2013) and is often used, although other authors have suggested that |r| can be set at a value between 0.6 and 0.9 (Randin et al., 2009; Young, 2010; Andreo et al., 2011; Syfert et al., 2013).

Predictive maps based on different predictor subsets

The comparison of model performance was conducted with a total of 440 MaxEnt models by using 22 predictor subsets with 20 replicates. The predictive maps of mean occurrence probability using 22 predictor subsets are presented in Supplementary Figure 1. Furthermore, Supplementary Figure 1 reveals that the predicted distribution area expanded when the prediction variables were R6, C6, or E6, which means that some unused variables may have still had predictive power. The overall patterns of 22 predictive maps were consistent, although some slight differences were observed among the maps. Supplementary Table 4 lists the spatial correlation coefficient (r) across 22 predictive maps calculated by pairing grid-based probabilistic values. The correlation coefficients of most pairs were more than 0.80, which indicated a high consistency among the 22 maps. Lower correlation coefficients (<0.79) were noted in the predictive maps that used R6 and E6 subsets, but all of these were greater than 0.67. Although these results mean that the three selection strategies are all effective, we still cannot know which strategy or environmental factor is the most ecologically significant. Therefore, the comprehensive application of these three strategies may be a better and more practical approach. For example, when the R strategy is used alone, LON will be selected

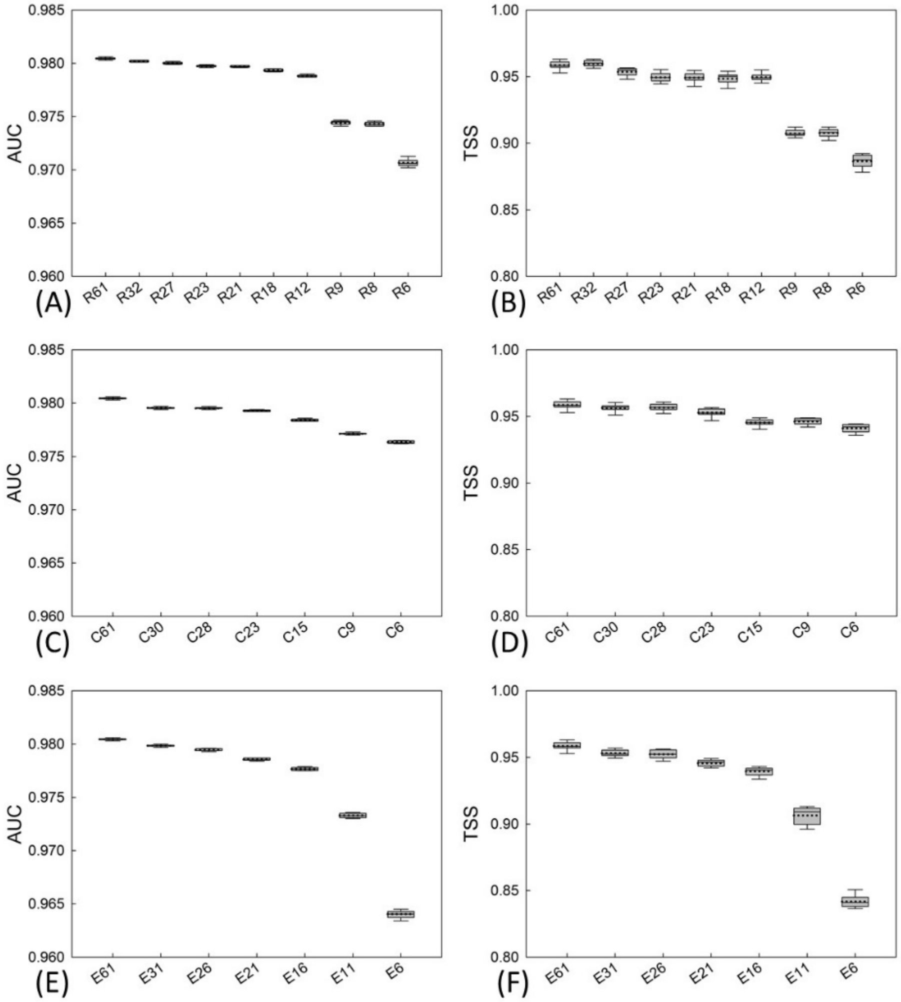

FIGURE 3 Comparison of model performance. Model performance was compared using the ROC by AUC and TSS criteria within the (A, B) 10-R subsets, (C, D) 7-C subsets, and (E, F) 7-E subsets. Each box-plot with a mean (dotted line) illustrates results of 20 cross-validation runs based on different predictor subsets by using backward elimination.

because of its low correlation coefficient with other factors, but if we use the comprehensive application of the three strategies, LON can be screened out in E strategy by experts who judge that its ecological significance is low.

For the visual comparison of predictive maps, the average and difference in predicted occurrence probability according to the three backward elimination procedures are illustrated in Figure 4. The average occurrence probability (mean ± standard deviation) of maps (left of Figure 4) modeled using 10-R subsets, 7-C subsets, and 7-E subsets was 0–0.8710 (0.0243 ± 0.133), 0–0.7288 (0.0187 ± 0.0038), and 0–0.8776 (0.0244 ± 0.0125), respectively. Therefore, the average probability modeled by predictor subsets from three procedures was highly similar (also see Supplementary Figure 1). The results implied that using backward elimination of redundant variables that considered the correlation coefficients between predictors (Dormann et al., 2013; Syfert et al., 2013), contribution level of predictors (Young, 2010; Yang et al., 2013), or expert selection of biologically relevant predictors (Watling et al., 2012; Harris et al., 2013) are all reasonable methods.

Differences in occurrence probability (right of Figure 4) of R61−R6, C61−C6, and E61−E6 were −0.8140 to 0.7651 (mean: −0.0263), −0.5297 to 0.6467 (mean: −0.0049), and −0.7619 to 0.8294 (mean: −0.0253), respectively. A detailed comparison of predictive maps is also presented in Supplementary Figure 1. Overall, the predictive probability modeled using a large number of predictors (R61, C61, and E61) was less than that of small number of predictors (R6, C6, and E6). This is possibly because the use of too many predictors in SDM results in over-parameterization and overfitting of the model (Guisan and Zimmermann, 2000; Tyberghein et al., 2012). The situation was true in our case with high collinearity or highly correlated predictors (see Supplementary Table 1), such as |r| > 0.9 between T1–T12, Bio1, Bio8, Bio9, Bio10, BT, WI, EWI, and ELE.

By the images in Figure 4, we can see that the difference between the full 61-variable model and the simplest 6-variable model is not unacceptable, it means that there is no need to use too many variables in SDM (Guisan and Zimmermann, 2000; Tyberghein et al., 2012).

Proposed approach to selecting predictors

Overall, the three procedures (high correlation, contribution level, and expert knowledge) for selecting predictors provided satisfactory results with high performance (Figures 3 and 4). However, the three procedures had their respective limitations. For example, biologically relevant predictors with strong correlations were vague in the selection procedure. Furthermore, regarding the contribution level of the procedures, collinearity was observed between predictors, such as Bio5 vs. T6 (|r| = 0.9), as presented in Table 3. Finally, the shortcomings of the expert knowledge procedure were not limited to the collinearity between predictors (Table 3) but also included artificial subjective decisions. Although numerous studies have been conducted on *F. hayatae* (Shen et al., 2015; Ying et al., 2016), identifying the most suitable set of environmental predictors for SDM for *F. hayatae* was difficult. Consequently, we propose a synthesis approach by combining the aforementioned three procedures to select predictors.

According to the aforementioned results, we suggest the following simple approach of combining different procedures of predictor selection.

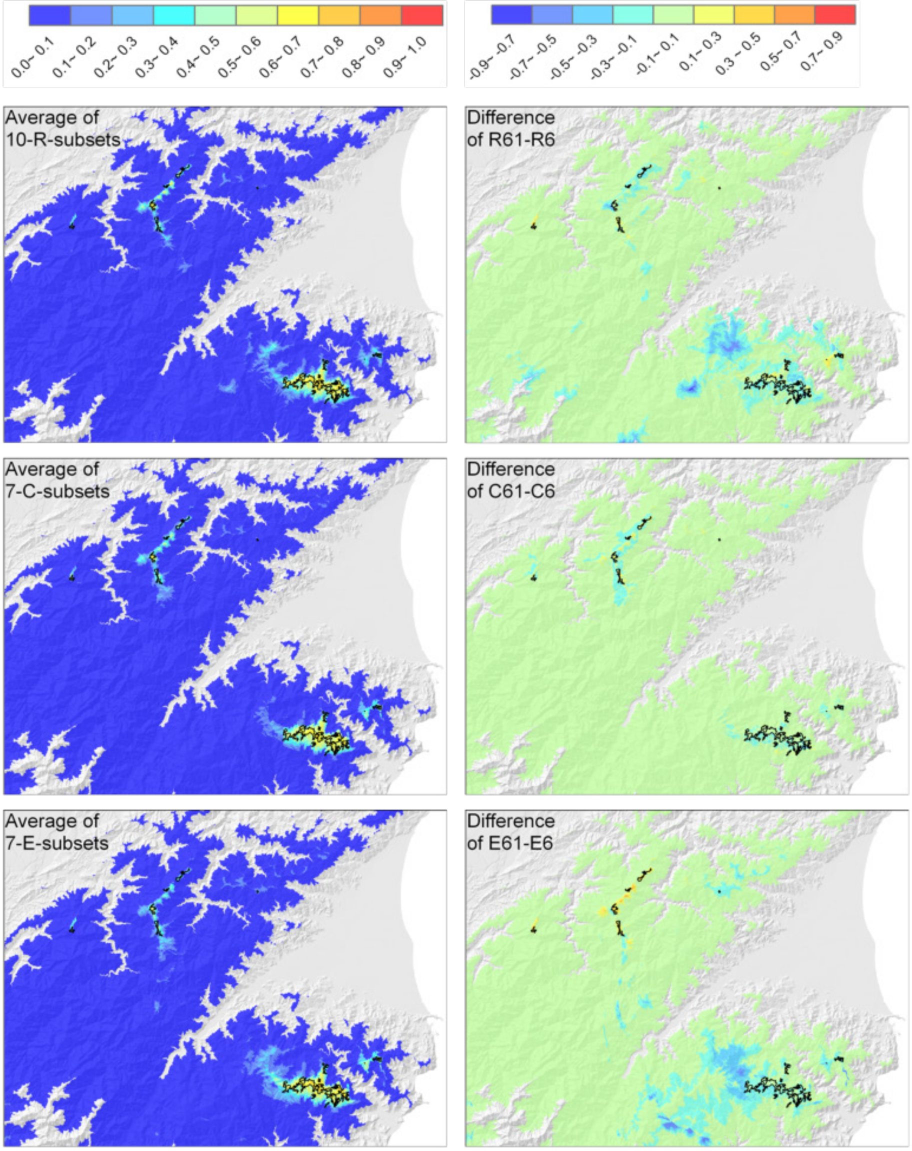

FIGURE 4 Average of and difference in probability maps. The average of (left) and difference in (right) probability maps based on different MaxEnt predictor subsets from three backward elimination procedures.

1. Identify and eliminate ineffective variables with nonidentifiability, such as CI in our case, through bivariate scatterplot analysis.

2. Calculate the correlation coefficients between other candidate predictors

3. Gradually select predictors within some highly correlated candidate subset on the basis of expert knowledge concerning biologically relevant predictors for target species and a low contribution level, which facilitates rejection of redundant predictors.

In order to understand whether the synthetic (abbreviated as S) strategy is superior to the other three (R, C, E) strategies, we choose 7 environmental factors for these four strategies to simulate the distribution of *F. hayatae*. Figure 5 compares SDM performance using R, C, E, and S strategies to select 7 environmental factors. The results show that S strategy is closer to the true distribution of *F. hayatae* than the other three strategies, and the S strategy is more clearly to reveal that the distribution of *F. hayatae* is mainly controlled by thermal-moisture regime and topographic location. Consequently, the synthetic selection strategy of environmental factors proposed in this paper can help to select the predictors suitable for SDM.

REFERENCES

ALLOUCHE, O.; TSOAR, A.; KADMON, R. Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). **Journal of Applied Ecology**, v. 43, n. 6, p. 1223-1232, 2006.

ANDREO, V.; GLASS, G.; SHIELDS, T.; PROVENSAL, C.; POLOP, J. Modeling potential distribution of *Oligoryzomys longicaudatus*, the Andes Virus (Genus: *Hantavirus*) reservoir, in Argentina. **EcoHealth**, v. 8, n. 3, p. 332-348, 2011.

ARAÚJO, M. B.; GUISAN, A. Five (or so) challenges for species distribution modelling. **Journal of Biogeography**, v. 33, n. 10, p. 1677-1688, 2006.

AUSTIN, M. P.; VAN NIEL, K. P. Improving species distribution models for climate change studies: Variable selection and scale. **Journal of Biogeography**, v. 38, n. 1, p. 1-8, 2011.

BEYER, H. L. 2012. **Geospatial modelling environment (version 0.7.2.0)**. Available at: http://www.spatialecology.com/gme. Accessed in: January 15th 2015.

CHIOU, C. R.; HSIEH, C. F.; WANG, J. C.; CHEN, M. Y.; LIU, H. Y.; YEH, C. L.; YANG, S. Z.; CHEN, T. Y.; HSIA, Y. J.; SONG, G. Z. M. The first national vegetation inventory in Taiwan. **Taiwan Journal of Forest Science**, v. 24, n. 4, p. 295-302, 2009.

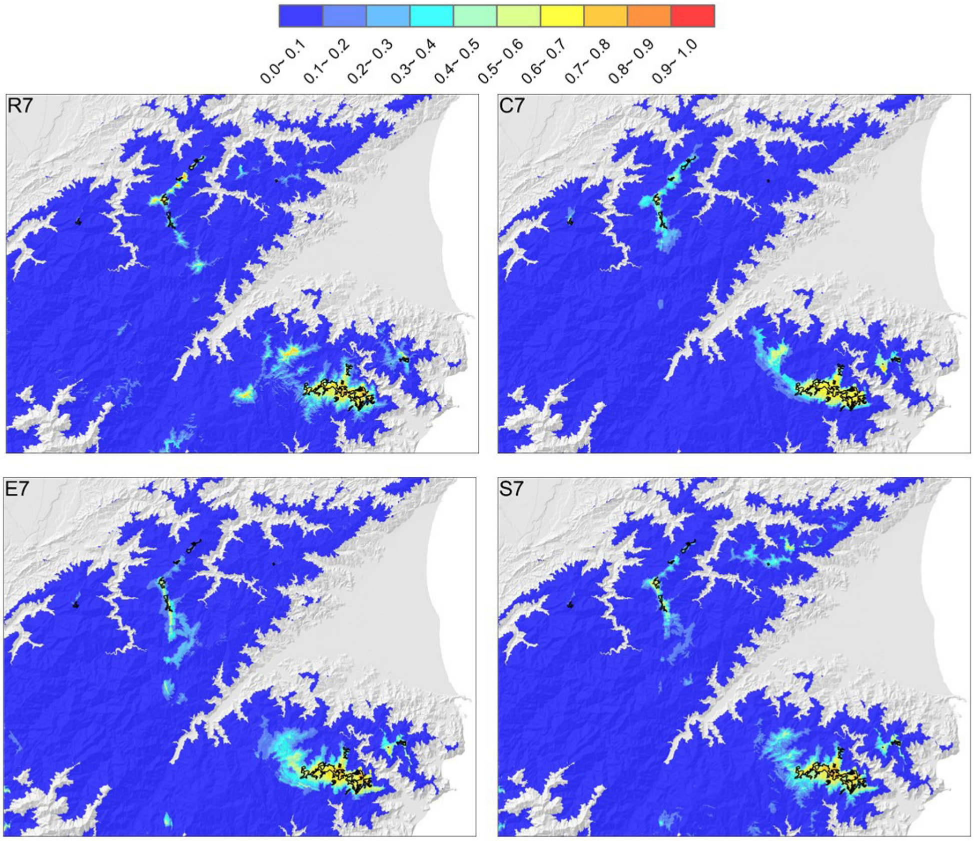

FIGURE 5 Comparison of *Fagus hayatae* predictive maps using R, C, E, and synthetic (abbreviated as S) strategies to select 7 factors. The area enclosed by the black line is the true distribution area of *F. hayatae*.

CHIU, C. A.; CHIOU, C. R.; LIN, J. R.; LIN, P. H.; LIN, C. T. Coldness index does not indicate the upper limit of evergreen broad-leaved forest in a subtropical island. **Journal of Forest Research**, v. 19, n. 1, p. 115-124, 2014.

CHIU, C. A.; LIN, P. H.; HSU, C. K.; SHEN, Z. H. A novel thermal index improves prediction of vegetation zones: Associating temperature sum with thermal seasonality. **Ecological Indicators**, v. 23, p. 668-674, 2012.

CHIU, C. A.; LIN, P. H.; LU, K. C. GIS-based tests for quality control of meteorological data and spatial interpolation of climatic data: A case study in mountainous Taiwan. **Mountain Research and Development**, v. 29, n. 4, p. 339-349, 2009.

CHIU, C. A.; LIN, P. H.; TSAI, C. Y. Spatio-temporal variation and monsoon effect on the temperature lapse rate of a subtropical island. **Terrestrial, Atmospheric and Oceanic Sciences**, v. 25, n. 2, p. 203-217, 2014.

DORMANN, C. F. Modelling species' distributions. In: JOPP, F.; REUTER, H.; BRECKLING, B. **Modelling complex ecological dynamics: An introduction into ecological modelling for students, Teachers & Scientists**. Springer, 2011. p. 179-196.

DORMANN, C. F.; ELITH, J.; BACHER, S.; BUCHMANN, C.; CARL, G.; CARRÉ, G.; GARCÍA MARQUÉZ, J. R.; GRUBER, B.; LAFOURCADE, B.; LEITÃO, P. J.; MÜNKEMÜLLER, T.; MCCLEAN, C.; OSBORNE, P. E.; REINEKING, B.; SCHRÖDER, B.; SKIDMORE, A. K.; ZURELL, D.; LAUTENBACH, S. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. **Ecography**, v. 36, n. 1, p. 27-46, 2013.

ELITH, J.; PHILLIPS, S. J.; HASTIE, T.; DUDÍK, M.; CHEE, Y. E.; YATES, C. J. A statistical explanation of MaxEnt for ecologists. **Diversity and Distributions**, v. 17, n. 1, p. 43-57, 2011.

EVANS, J. 2011. **Geomorphometry and gradient metrics toolbox**. Available at: http://conserveonline.org/workspaces/emt/documents/arcgis-geomorphometrics-toolbox/view.html. Accessed in: August 8th 2014.

FRANKLIN, J. **Mapping species distributions: Spatial inference and prediction**. Cambridge University Press, 2010. 340 p.

GONTIER, M.; MÖRTBERG, U.; BALFORS, B. Comparing GIS-based habitat models for applications in EIA and SEA. **Environmental Impact Assessment Review**, v. 30, n. 1, p. 8-18, 2010.

GUISAN, A.; THUILLER, W. Predicting species distribution: Offering more than simple habitat models. **Ecology Letters**, v. 8, n. 9, p. 993-1009, 2005

GUISAN, A.; ZIMMERMANN, N. E. Predictive habitat distribution models in ecology. **Ecological Modelling**, v. 135, n. 2-3, p. 147-186, 2000.

GUISAN, A.; ZIMMERMANN, N. E.; ELITH, J.; GRAHAM, C. H.; PHILLIPS, S.; PETERSON, A. T. What matters for predicting the occurrences of trees: Techniques, data, or species' characteristics? **Ecological Monographs**, v. 77, n. 4, p. 615-630, 2007.

HARRIS, R.; PORFIRIO, L. L.; HUGH, S.; LEE, G.; BINDOFF, N. L.; MACKEY, B.; BEETON, N.J. To be or not to be? Variable selection can change the projected fate of a threatened species under future climate. **Ecological Management & Restoration**, v. 14, n. 3, p. 230-234, 2013.

HIJMANS, R. J.; CAMERON, S. E.; PARRA, J. L.; JONES, P. G.; JARVIS, A. Very high resolution interpolated climate surfaces for global land areas. **International Journal of Climatology**, v. 25, n. 15, p. 1965-1978, 2005.

HOLDRIDGE, L. R. **Life zone ecology**. Tropical Science Center, 1967. 148 p.

KIRA, T. Forest ecosystems of east and southeast Asia in a global perspective. **Ecological Research**, v. 6, n. 2, p. 185-200, 1991.

LAI, Y. J.; CHOU, M. D.; LIN, P. H. Parameterization of topographic effect on surface solar radiation. **Journal of Geophysical Research**, v. 115, n. D1, p. D01104, 2010.

LIPPITT, C. D.; ROGAN, J.; TOLEDANO, J.; SANGERMANO, F.; EASTMAN, J. R.; MASTRO, V.; SAWYER, A. Incorporating anthropogenic variables into a species distribution model to map gypsy moth risk. **Ecological Modelling**, v. 210, n. 3. p. 339-350, 2008.

LIU, C.; WHITE, M.; NEWELL, G. Selecting thresholds for the prediction of species occurrence with presence-only data. **Journal of Biogeography**, v. 40, n. 4, p. 78-789, 2013.

MAC NALLY, R. Multiple regression and inference in ecology and conservation biology: Further comments on identifying important predictor variables. **Biodiversity & Conservation**, v. 11, n. 8, p. 1397-1401, 2002.

MORISETTE, J. T.; JARNEVICH, C. S.; HOLCOMBE, T. R.; TALBERT, C. B.; IGNIZIO, D.; TALBERT, M. K.; SILVA, C.; KOOP, D.; SWANSON, A.; YOUNG, N. E. VisTrails SAHM: Visualization and workflow management for species habitat modeling. **Ecography**, v. 36, n. 2, p. 129-135, 2013.

PETITPIERRE, B.; BROENNIMANN, O.; KUEFFER, C.; DAEHLER, C.; GUISAN, A. Selecting predictors to maximize the transferability of species distribution models: Lessons from cross-continental plant invasions. **Global Ecology and Biogeography**, v. 26, n. 3, p. 275-287, 2017.

PORFIRIO, L. L.; HARRIS, R. M.; LEFROY, E. C.; HUGH, S.; GOULD, S. F.; LEE, G.; BINDOFF, N. L.; MACKEY, B. Improving the use of species distribution models in conservation planning and management under climate change. **PLoS ONE**, v. 9, n. 11, p. e113749, 2014.

QUINN, G. G. P.; KEOUGH, M. J. **Experimental design and data analysis for biologists**. Cambridge University Press, 2002. 537 p.

RANDIN, C. F.; VUISSOZ, G.; LISTON, G. E.; VITTOZ, P.; GUISAN, A. Introduction of snow and geomorphic disturbance variables into predictive models of alpine plant distribution in the Western Swiss Alps. **Arctic, Antarctic, and Alpine Research**, v. 41, n. 3, p. 347-361, 2009.

SANGERMANO, F.; EASTMAN, J. R. A GIS framework for the refinement of species geographic ranges. **International Journal of Geographical Information Science**, v. 26, n. 1, p. 39-55, 2012.

SHEN, Z. H.; FANG, J. F.; CHIU, C. A.; CHEN, T. Y. The geographical distribution and differentiation of Chinese beech forests and the association with *Quercus*. **Applied Vegetation Science**, v. 18, n. 1, p. 23-33, 2015.

SOLOMON, B. S.; DUGGAN, A. K.; WEBSTER, D.; SERWINT, J. R. Pediatric residents' attitudes and behaviors related to counseling adolescents and their parents about firearm safety. **Archives of Pediatrics & Adolescent Medicine**, v. 156, n. 8, p. 769-775, 2002.

SU, H. J. Studies on the climate and vegetation types of the natural forests in Taiwan (3): A scheme of geographical climatic regions. **Quarterly Journal of Chinese Forestry**, v. 18, n. 3, p. 33-44, 1985.

SWETS, J. A. Measuring the accuracy of diagnostic systems. **Science**, v. 240, n. 4857, p. 1285-1293, 1988.

SYFERT, M. M.; SMITH, M. J.; COOMES, D. A. The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. **PLoS ONE**, v. 8, n. 2, p. e55158, 2013.

TYBERGHEIN, L.; VERBRUGGEN, H.; PAULY, K.; TROUPIN, C.; MINEUR, F.; DE CLERCK, O. Bio-ORACLE: A global environmental dataset for marine species distribution modelling. **Global Ecology and Biogeography**, v. 21, n. 2, p. 272-281, 2012.

WATLING, J. I.; ROMAÑACH, S. S.; BUCKLIN, D. N.; SPEROTERRA, C.; BRANDT, L. A.; PEARLSTINE, L. G.; MAZZOTTI, F. J. Do bioclimate variables improve performance of climate envelope models? **Ecological Modelling**, v. 246, p. 79-85, 2012.

WILLIAMS, K. J.; BELBIN, L.; AUSTIN, M. P.; STEIN, J. L.; FERRIER, S. Which environmental variables should I use in my biodiversity model? **International Journal of Geographical Information Science**, v. 26, n. 11, p. 2009-2047, 2012.

WISZ, M. S.; HIJMANS, R. J.; LI, J.; PETERSON, A. T.; GRAHAM, C. H.; GUISAN, A. Effects of sample size on the performance of species distribution models. **Diversity and Distributions**, v. 14, n. 5, p. 63-773, 2008.

WOLFE, J. A. Temperature parameters of humid to mesic forest of eastern Asia and relation to forests of other regions of the northern hemisphere and Australasia. **Geological Survey Professional Paper**, v. 1106, p. 1-37, 1979.

XU, W. D. The application of Kira's thermal index to Chinese vegetation. **Chinese Journal of Ecology**, v. 4, n. 3, p. 35-39, 1985. [in Chinese with English abstract]

YANG, X. Q.; KUSHWAHA, S. P. S.; SARAN, S.; XU, J.; ROY, P. S. Maxent modeling for predicting the potential distribution of medicinal plant, Justicia adhatoda L. in Lesser Himalayan foothills. **Ecological engineering**, v. 51, p. 83-87, 2013.

YING, L. X.; ZHANG, T. T.; CHIU, C. A.; CHEN, T. Y.; LUO, S. J.; CHEN, X. Y.; SHEN, Z. H. The phylogeography of *Fagus hayatae* (Fagaceae): Genetic isolation among populations. **Ecology and Evolution**, v. 6, n. 9, p. 2805-2816, 2016.

YOUNG, N. E. **Regional data refine local abundance models: Modeling plant species abundance distributions on the central plains**. 2010. 43 p. Thesis, Department of Forest, Rangeland, and Watershed Stewardship, Colorado State University, Colorado, USA.

ZHANG, Z.; XU, S.; CAPINHA, C.; WETERINGS, R.; GAO, T. Using species distribution model to predict the impact of climate change on the potential distribution of Japanese whiting *Sillago japonica*. **Ecological Indicators**, v. 104, p. 333-340, 2019.