Scientific Research Publishing

Distribution Prediction Model of a Rare Orchid Species (*Vanda bicolor* Griff.) Using Small Sample Size

Chitta Ranjan Deb, N. S. Jamir, Zubenthung P. Kikon

Department of Botany, Nagaland University, Nagaland, India
Email: debchitta@rediffmail.com and debchitta@gmail.com

Abstract

Advancement in field of GIS and Information Technology has taken conservation works and strategies a step further as most conservation works are now dependent on these technologies. The present study explores the prediction ability of MAXENT using a very low sample size by applying jackknife analysis over a well defined smaller region and using only climate data. *Vanda bicolor* is a horticulture important orchid grown in certain patches of North Eastern region of India and the species considered to be "Vulnerable". Present study reports a distribution prediction model using different geo-climatic parameters for a small area. Model validation by ground truthing gives a significant successful result which clearly defines the ability of MAXENT prediction model to give high success rate (71%) with low training samples. Use of the low sample size over a larger area results in unstable models however application of these samples in smaller radius around the occurrence points could provide good working models.

Keywords

MAXENT Distribution Prediction Model, ENM, Ground Truthing, *Vanda bicolor*

1. Introduction

Recent change in climatic condition has increased the pressure on plant species and many important species has been subjected to a lot of stress pushing them to the brink of extinction. In this aspect the issue of conservation has become a topic of utmost importance. Many plant species has been destroyed before they are documented or their value is realized, most of the medicinal and ornamental plants are highly exploited for their economic value. In this scenario knowledge

of their distribution and the niche radius of target plant species will allow conservators to work out conservation strategies effectively for those species that are at risk. Species distribution modeling has become an important tool for conservation works as it provides an insight to the species geographical and climatic requisites and this data can be of immense help for conservationist.

For development of niche model large numbers of climatic variables are required. However, there are certain species especially rare and threatened species their distribution is restricted in narrow geographical area and not possible to have large numbers of variables to develop a robust distribution model. In the recent past there are few reports published on developing models using low sample size giving significant results e.g., sample size of 2 - 4 [1] [2] [3], sample size of 16 - 17 [4] [5]. Further in the recent past prediction models has been developed using low sample size with as few as 5 - 10 observations [6]. *Vanda bicolor* is a horticultural important orchid found in North Eastern region of India especially in Nagaland. The species is considered to be vulnerable and demands efforts for conservation. Present study was undertaken to develop a distribution prediction model and climate suitability model for conservation of the species. Present study was aimed to bring out the environmental data within which the target species can persist and the geographical area where these set of environmental conditions are offered. Those areas offering similar climatic set as that of the training site are potential sites for their reintroduction taking into consideration the land-use pattern and biotic interaction in defining the species prevalence in the predicted regions as defined by Pearson and Dawson [7]. Past studies have shown that the predictive performance decreases significantly when samples size are low [3] and in our study we investigated the performance of the models developed using low sample size using *Vanda bicolor* a rare orchid species as the target species. To enable the assessment of the prediction ability of the model developed using small sample size we employed the Jackknife of Maximum Entropy (MAXENT) [8].

2. Materials and Methods

2.1. Target Species

Vanda bicolor Griff. is a rare epiphytic orchid belonging to the genus *Vanda* under family-Orchidaceae, the plant has leafy stem enveloped almost fully covered by leaf sheaths and each with oblong, curved, and with little twisting in the middle, apically obliquely bilobed, each lobe tridentate leaves. *Vanda bicolor* flowers between March to June and inflorescence axillary, glabrous flowers white-purplish, mottled above, violet tinged beneath, with floral size of 4 - 6 cm (Figure 1). The plant is found mostly in Tropical Wet evergreen forest, tropical Semi Evergreen forest and Sub-tropical broad leaved wet hill forest, This orchid has an endemic distribution restricted to Indo-Burma regions of India Arunachal Pradesh, Assam, Nagaland, Sikkim; Bhutan, Myanmar, Nepal [9] [10] [11] [12].

Figure 1. *Vanda bicolor* in bloom.

2.2. Study Site

The study was carried out in Nagaland, India which lies between 93˚20' - 95˚15' East Longitude and 25˚31'- 27˚1' North Latitude with a total geographical area of 16,579 km². The state has a forest cover of 13,044 km² (78.68%) according to Forest Survey of India, 2013 with a total forest cover loss of 274 Km² since 2011 report. The state falls under seven forest types *viz.* Tropical Wet Evergreen, Tropical Semi Evergreen, Tropical Moist Deciduous, Subtropical Broad Leaved Hill, subtropical pine and Montane Temperate Forests with an average rainfall of 1583 mm [9]. Nagaland has a monsoon type climate and enjoys a salubrious climate. Annual rainfall ranges around 70 - 100 inches (1800 - 2500 mm), in the months between May to September. Temperatures range from 70˚F (21˚C) to 104˚F (40˚C). In winter, temperatures do not generally drop below 39˚F (4˚C), but frost is common at high elevations. Summer is short lasting only for a few months with temperature between 16˚C (61˚F) to 31˚C (88˚F). The state experiences an early winter which is cold and dry, The maximum average temperature recorded in the winter season is 24˚C (75˚F). (http://www.nagenvis.nic.in/Database/Climate_884.aspx).

2.3. Input Data

In the present study only 4 occurrence points were used to develop the model, all presence points are geo-referenced from primary ground surveys using GPS. The occurrence points are subjected to quality test with respect to and their positional accuracy was ascertained through Google earth, duplicates are identified and removed thus maintaining only one point within 1×1 Km² to avoid sampling bias which would otherwise favor the climatic of those sites where sampling is highly concentrated. As the number of presence points in below 20 1.5 × Interquartile range (1.5 IQR) method of identifying outliers is applied to check

for outliers based on climate data developed from the environmental data obtained from WorldClim Website at 30''. All climate data are cross checked for resolution accuracy and corrected to 30'' pixel resolution.

19 bioclimatic variables of zone 29 was obtained from WorldClim at 30'' pixel resolution, which consist of an interpolated datasets of temperature and precipitation [10], which are of primary importance for the plant to thrive and reproduce successfully at a particular area. The basic idea was to identify the most important variable and the month influencing the model.

2.4. Ecological Niche Modeling (ENM)

All modeling works was carried out using MAXENT Version 3.3.3 K as our works are based on presence points only and has low sample size [6]. MAXENT is designed to efficiently handle small sample size; all visualization was done in DIVA GIS 7.5.0, whereas all mapping works was carried out using ARC GIS 9.3. The model was developed using Jacknifing method [3], For validating model robustness, 12 replicated model runs was executed with a threshold rule of 10 percentile training presence and employed cross validation technique for dividing the samples into replicate folds and using as test data all other parameters were kept at default. The Area under Curve (AUC) was graded according to [11]. The distribution potential of the model was classified into very low (0 - 0.2), low (0 - 0.4), medium (0 - 0.6), high (0 - 0.8) and very high (0 - 1) and basing on the prediction model the target plant was introduced in all the different prediction threshold to study the response of the plant in the different predicted areas.

3. Results

The model calibration gives a test AUC of 0.984, with a standard deviation of 0.004. The AUC ranges from 0.5 for models that are no better than random to 1.0 for models with perfect predictive ability (Table 1, Figure 2). The AUC test is derived from the Receiver Operating Characteristic (ROC) Curve. The ROC curve thus describes the relationship between the proportion of observed presences correctly predicted (sensitivity) and the proportion of observed absences incorrectly predicted (1–specificity). Thus, an AUC value of 0.7 means the probability is 0.7 that a record selected at random from the set of presences will have a predicted value greater than a record selected at random from the set of absences [12] [13]. In the present study estimates of relative contributions of the environmental variables to the MAXENT model showed that Bio18 contributed the maximum (46.7%) followed by Bio13 and contributing 16.2 and 9.4% (Table 1).

MAXENT jackknife test of variable (Figure 2) importance shows Bio18 (Precipitation of Warmest Quarter) giving a reasonably good fit to the training data. The environmental variable with highest gain when used in isolation is Bio18, which therefore appears to have the most useful information by itself in the model. The environmental variable that decreases the gain the most when it is
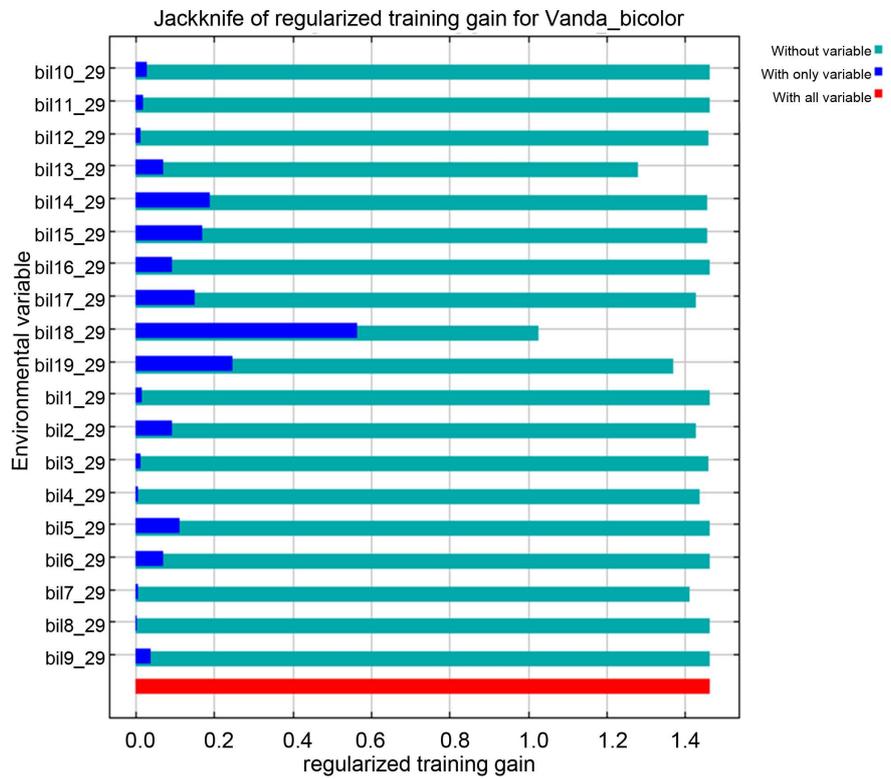
Figure 2. Result for the jackknife test of variable importance.

Table 1. Contributions of the environmental variables to the MAXENT models using the 19 Bioclimatic variables.

Variable	% contribution	Permutation importance (%)
bio18	50.3	31.6
bio13	13.8	1.9
bio19	11.5	23.3
bio2	6.8	4.7
bio14	6.3	5.6
bio7	4.5	18.8
bio5	3.4	0.2
bio17	0.9	7.4
bio4	0.8	2

omitted was also observed in bio18, which therefore appears to have the most information that isn't present in the other variables.

Same jackknife test, using test gain instead of training gain (Figure 3) also shows that Precipitation of Warmest Quarter as an important climatic variable in the test gain, the test gain plot also shows that a model made only using Bio8 (Mean temperature of wettest quarter) results in a negative test gain. The model thus is below a null model (*i.e.*, a uniform distribution) for predicting the distribution of occurrences set aside for testing and the variables are not the useful as predictor.
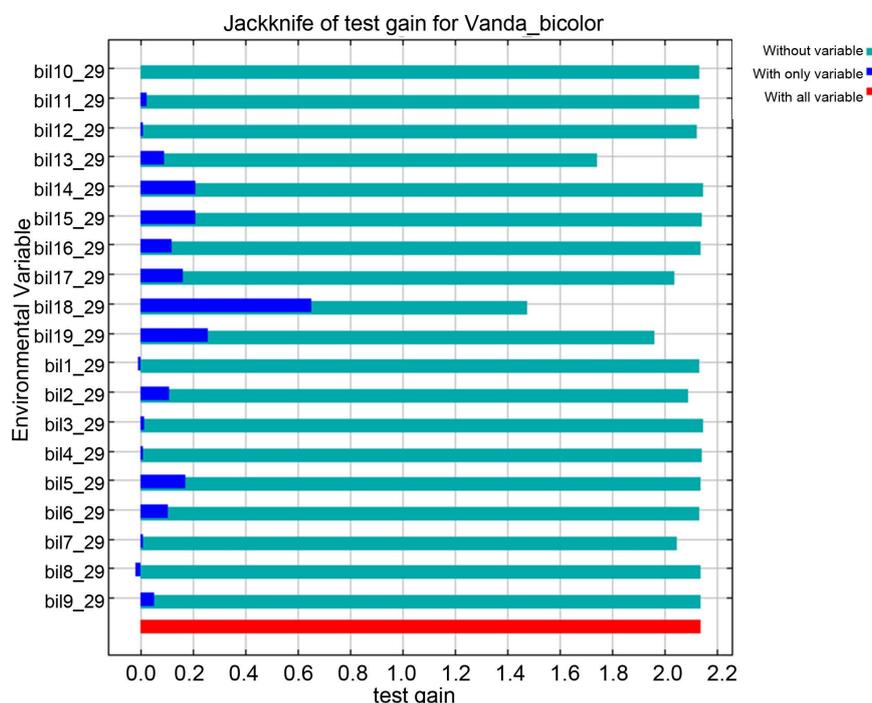
Figure 3. Result for the Jackknife test using test gain.

Jackknife test using AUC on test data, the AUC plot shows that Bio18 is the most effective single variable for predicting the distribution of the occurrence data that was left aside for testing, when the predictive performance is measured using AUC, though it was hardly used by the model built using all variables and the relative importance of Bio4 also increases in the test gain plot (**Figure 4**). These results establishes the importance of precipitation in the MAXENT prediction model and the role of model development for MAXENT to obtain a good fit to the training data with the Precipitation of Warmest Quarter defining better results on the set-aside test data (most useful variable as predictor) followed by BIO5 (Max Temperature of Warmest Period), BIO16 (Precipitation of Wettest Quarter) and BIO13 (Precipitation of Wettest Period).

3.1. Evaluating Model Validation and Authentication

The model was developed using a very low occurrence points and most of the areas of Nagaland was predicted under high suitability threshold, thus to validate this, the model was subjected to intensive ground truthing and introduction in different prediction threshold to assess the model prediction ability (**Figure 5**). Target plant species was introduced in different areas and ground truthing works was carried out and it was realized that the distribution of the target plant species was highly threatened and very sparsely distributed in pockets. *K*-fold partitioning of test data and training data could not give usable model as the occurrence points are too low therefore Pearsons jackknife method of leaving one out and assessing the predictive performance of each separate model was used. Jackknifing method was able to construct a workable model and ground
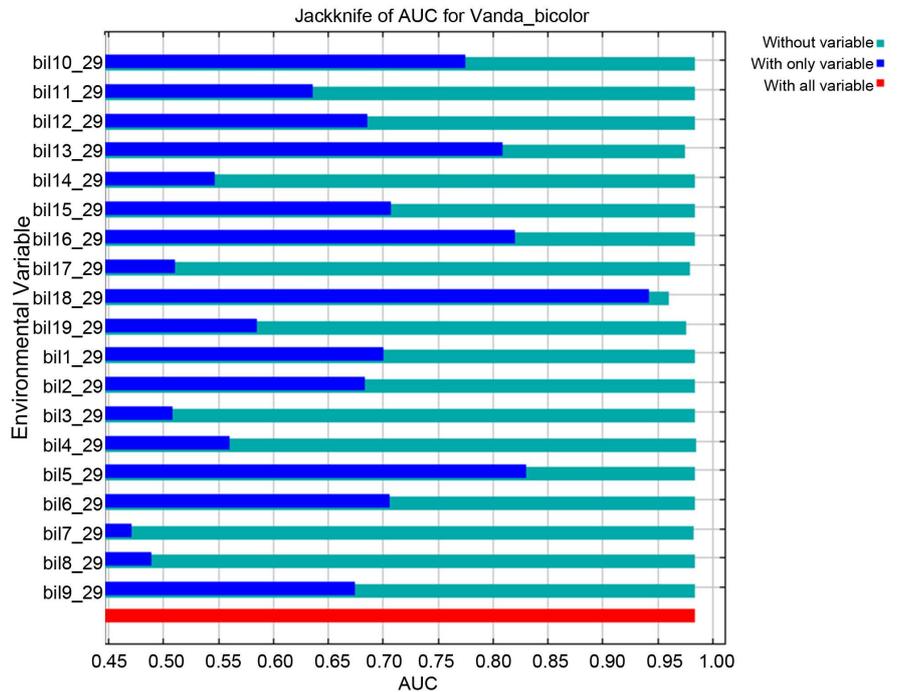
Figure 4. Result for the Jackknife test using AUC on test data.
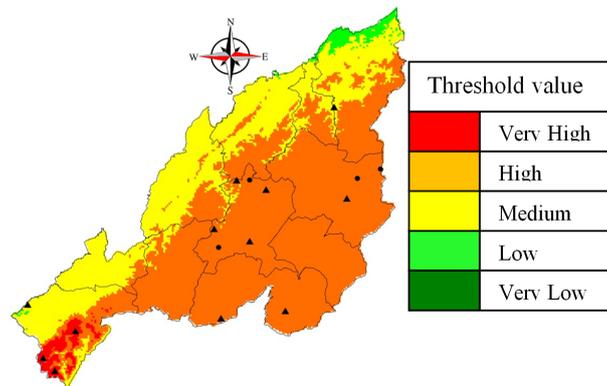

Figure 5. Prediction map of *Vanda bicolor* in Nagaland. (●-Training Sample, ▲-Newly discovered occurrence).

truthing works by random selection of sites in different prediction threshold level give a significant result with 10 new occurrence records (Table 2). The MAXENT model was able to give significant prediction results over a smaller area however, when the small sample size data was used to predict over a larger area *i.e.*, whole part of India and North Eastern region of India, the predication model becomes unstable and insignificant.

Though the model was developed using only four training sites, it was able to predict suitable sites in the neighboring Northeastern states of India and countries (Figure 6). The high suitability threshold was validated in Manipur and Arunachal, with secondary occurrence data, the model prediction in neighboring countries of Bhutan and Burma can also be supported by occurrence reports available from secondary sources. The ability of the model to predict all the
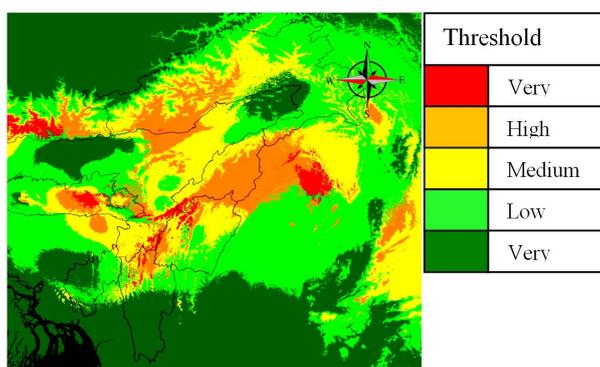
Figure 6. Prediction map of *Vanda Bicolor* showing prediction in the neighboring states of the target area

Table 2. Occurrence data in Nagaland acquired form ground truthing.

Area	Longitude	Latitude	Prediction Threshold	Niche Status
Izheto	94.4200	26.2041	High	Disturbed forest, frequent Jhuming
Ghukimi	94.3110	25.9683	High	Disturbed forest, frequent Jhuming
Tsupfume	94.3443	25.5342	High	Disturbed forest, frequent Jhuming and forest fire
Aopao	94.8919	26.5591	Medium	Disturbed forest, frequent Jhuming
Ghokhuye	94.4832	25.9083	High	Disturbed forest, frequent Jhuming
Kengnyu	94.9535	26.1159	High	Disturbed forest, frequent Jhuming
Reguri	94.6557	25.5701	High	Disturbed forest, frequent Jhuming
Chisailhem	93.4831	25.3424	Very high	Disturbed forest, frequent Jhuming and forest fire
Nsong	93.5392	25.2798	Very high	Disturbed forest, frequent Jhuming and forest fire
Old Tesen	93.6391	25.473	Very high	Disturbed forest, frequent Jhuming and forest fire

suitable sites over larger areas might be lowered as the training points are very less and confined over a smaller area (*i.e.*, Nagaland), The model however shows a more robust prediction outside the target area in Bhutan and Myanmar.

3.2. Conservation Planning and Prioritizing Areas for Reintroduction

During the present study it was observed that most of the occurrence areas are under high biological disturbances like logging, Jhuming and forest fire and these are some of the factors that are bringing noticeable changes to the forest over a short period. This spatially separated population shares similarity in host plant and seasonal climatic variables like precipitation and temperature. Most of the areas are under high suitability threshold but are under high anthropogenic disturbances and only a small portion of the study in very high suitability threshold a falls under undisturbed area and interestingly Intanki National Park, In-

dia fall under very high suitability threshold. Introduction of species to random forests will proved to be futile if careful assessment of the forest condition is not done areas like Intanki national park will serve as excellent sites for in-situ conservation and possible re-introduction for species recovery.

4. Discussion

The study was able to produce significant prediction models using very small sample size over a defined area, which has been validated statistically and though ground truthing. Earlier studies on development of models using low sample size has also reported effective models by using sample size of minimum 4 and 5 study on cryptic geckos [3]. A testing of four modeling methods (Bioclim, Domain, GARP, and MAXENT) across 18 species with different levels of ecological specialization using six different sample size treatments and three different evaluation measures revealed that MAXENT was the most capable of the four modeling methods in producing useful results with sample sizes as small as 5, 10 and 25 occurrences. MAXENT also predicted the largest area of all modeling methods at sample size 5 and remained fairly level at sample sizes of 25 and above [6]. In the present study MAXENT was used to develop the model using very low sample size over a smaller area of 16,579 km² to predict the climatic suitability of the target species. The model was able to bring out interesting insights on the climatic parameters which are playing an important role in the survival of the target species; the model defining precipitation as the most important predictor for the model. During field survey it was observed that occurrence areas are mostly hot and humid with high rainfall and the possibility of precipitation playing an important role in maintaining the population of the target species is quite relevant. The basic target for conservation works are on those species that are under high threat and those species in high threat category usually have low occurrence and it will be insignificant for conservationist unless working models are developed for these threatened species. In this study only climate data was used as our target species is plant species and climate plays a major role for its well being however, this does not negates the possibilities of anthropogenic and other biological factors contributing to the habitat loss and low occurrence of species. The predication ability of the model with low sample size over a smaller area can be used to develop a mosaic of prediction models in areas where occurrence points are less and are in considerably distant pockets. Our study was able to give a success rate of 70% (calculated on stack developed using the MAXENT prediction map threshold value over the area of occurrence) with just 4 sample over a small well defined area. Introduction of the target plants to high and medium prediction threshold area (Wokha-94˚15'29.72''E, 26˚05'47.10''E; Prediction Threshold-High; Mokokchung-94˚ 31'22.25''E, 26˚19'34.72''N; Prediction Threshold-High; Dimapur-93˚43'13.00''E, 25˚55'13.56''N; Prediction Threshold-Medium) gives normal flowering and seed formation. During ground truthing works it was found that most the sites of occurrences fall in areas where there is frequent Jhuming and forest fire, the practice of Jhum cultivation and

leaving it fellow for the next 5-10 years allows the regeneration of vegetation and the cycle is repeated and the survival rate of economically important plant species is much reduced. The present study thus could serve as a model for rehabilitations particularly for the conservation of economically important and rare plant species and gives ample scope to future researchers for more holistic studies on this approach and development of models with low training samples.

5. Conclusion

In the present study, the effectiveness of low sample size and climate data on MAXENT model development and its usability in real world application has been validated statistically through ground truthing and testing of sites by introducing plants to predicted sites. The model was able to bring out new insights on the climatic parameters which defines species survival, and successfully predicted new pollution in wild and those existing population in neighboring states and countries with success rate of 70% (calculated on stack developed using the MAXENT prediction map threshold value over the area of occurrence). Any conservation related works will be on those species that are under high threat and those species in high threat category usually have low occurrence and it will be insignificant for conservationist unless working models are developed for these threatened species and the present study gives a good example of how low sample size can be used to develop effective prediction models.

Acknowledgements

References

[1] Loiselle, B.A., Howell, C.A., Graham, C.H., Goerck, J.M., Brooks, T., Smith, K.G. and Williams, P.H. (2003) Avoiding Pitfalls of Using Species Distribution Models in Conservation Planning. *Conservation Biology,* **17**, 1591-1600. https://doi.org/10.1111/j.1523-1739.2003.00233.x

[2] Ortega-Huerta, M.A. and Peterson, A.T. (2004) Modeling Spatial Patterns of Biodiversity for Conservation Prioritization in North-Eastern Mexico. *Diversity and Distributions*, **10**, 39-54. https://doi.org/10.1111/j.1472-4642.2004.00051.x

[3] Pearson, R.G., Raxworthy, C.J., Nakamura, M. and Peterson, A.T. (2007) Predicting Species Distributions from Small Numbers of Occurrence Records: A Test Case Using Cryptic Geckos in Madagascar. *Journal of Biogeography*, **34**, 102-117. https://doi.org/10.1111/j.1365-2699.2006.01594.x

[4] Adhikari, D., Barik, S.K. and Upadhaya, K. (2012) Habitat Distribution Modelling for Reintroduction of *Ilex khasiana* Purk. A Critically Endangered Tree Species of Northeastern India. *Ecology Engineering*, **40**, 37-43.

[5] Groff, L.A., Marks, S.B. and Hayes, M.P. (2014) Using Ecological Niche Models to Direct Rare Amphibian Surveys: A Case Study Using the Oregon Spotted Frog (*Rana Pretiosa*). *Herpetological Conservation and Biology,* **9**, 354-368.

[6] Hernandez, P.A., Graham, C.H., Master, L.L. and Albert, D.L. (2006) The Effect of Sample Size and Species Characteristics on Performance of Different Species Distribution Modeling Methods. *Ecography,* **29**, 773-785.
https://doi.org/10.1111/j.0906-7590.2006.04700.x

[7] Pearson, R.G. and Dawson, T.P. (2003) Predicting the Impacts of Climate Change on the Distributions of Species: Are Bioclimate Envelope Models Useful? *Global Ecology and Biogeography,* **12**, 361-371.
https://doi.org/10.1046/j.1466-822X.2003.00042.x

[8] Phillips, S.J., Anderson, R.P. and Schapire, R.E. (2006) Maximum Entropy Modeling of Species Geographic Distributions. *Ecological Modelling,* **190**, 231-259.

[9] Noltie, H.J. (1994) Flora of Bhutan Including a Record of Plants from Sikkim and Darjeeling. Royal Botanical Garden Edinburgh, Inverleith Row, Edinburgh EH3 5LR, UK, **3**, 21-23.

[10] Hynniewta, T.M., Kataki, S.K. and Wadhwa, B.M. (2000) Orchids of Nagaland. Botanical Survey of India, Kolkata.

[11] Pearce, N.R. and Cribb, P.J. (2002) The Orchids of Bhutan. Royal Botanic Garden Edinburgh, 20A Inverleith Row Edinburgh EH3 5LR, and Royal Government of Bhutan.

[12] De, A. and Hajra, P.K. (2004) *Vanda* in India. *Journal of Orchid Society of India,* **18,** 28-29.

[13] Kurse, B.C. and Singh, Kh.S. (2012) Study of Spatial and Temporal Distribution of Rainfall in Nagaland (India). *International Journal of Geomatics and Geosciences,* **2**, 712-722.

Scientific Research Publishing