Journal of Threatened Taxa

Building evidence for conservation globally

www.threatenedtaxa.org
ISSN 0974-7907 (Online) | ISSN 0974-7893 (Print)

COMMUNICATION

EVALUATING PERFORMANCE OF FOUR SPECIES DISTRIBUTION MODELS USING BLUE-TAILED GREEN DARNER ANAX GUTTATUS (INSECTA: ODONATA) AS MODEL ORGANISM FROM THE GANGETIC RIPARIAN ZONE

Kritish De, S. Zeeshan Ali, Niladri Dasgupta, Virendra Prasad Uniyal, Jeyaraj Antony Johnson & Syed Ainul Hussain

26 October 2020 | Vol. 12 | No. 14 | Pages: 16962–16970

The Journal of Threatened Taxa (JoTT) is dedicated to building evidence for conservation globally by publishing peer-reviewed articles online every month at a reasonably rapid rate at www.threatenedtaxa.org. All articles published in JoTT are registered under Creative Commons Attribution 4.0 International License unless otherwise mentioned. JoTT allows unrestricted use, reproduction, and distribution of articles in any medium by providing adequate credit to the author(s) and the source of publication.

The opinions expressed by the authors do not reflect the views of the Journal of Threatened Taxa, Wildlife Information Liaison Development Society, Zoo Outreach Organization, or any of the partners. The journal, the publisher, the host, and the partners are not responsible for the accuracy of the political boundaries shown in the maps by the authors.
Evaluating performance of four species distribution models using Blue-tailed Green Darner *Anax guttatus* (Insecta: Odonata) as model organism from the Gangetic riparian zone

Kritish De¹, S. Zeeshan Ali², Niladri Dasgupta³, Virendra Prasad Uniyal¹, Jeyaraj Antony Johnson³ & Syed Ainul Hussain⁴

¹- Wildlife Institute of India, Chandrabani, Dehradun, Uttarakhand 248001, India.  
²kritish.de@gmail.com (corresponding author), ³zeeshanearth@gmail.com, ⁴niladri4all@gmail.com, ⁵uniyalvp@wii.gov.in, ⁶hussain@wii.gov.in

Abstract: In this paper we evaluated the performance of four species distribution models: generalized linear (GLM), maximum entropy (MAXENT), random forest (RF) and support vector machines (SVM) model, using the distribution of the dragonfly Blue-tailed Green Darner *Anax guttatus* in the Gangetic riparian zone between Bijnor and Kanpur barrage, Uttar Pradesh, India. We used forest cover type, land use, land cover and five bioclimatic variable layers: annual mean temperature, isothermality, temperature seasonality, mean temperature of driest quarter, and precipitation seasonality to build the models. We found that the GLM generated the highest values for AUC, Kappa statistic, TSS, specificity and sensitivity, and the lowest values for omission error and commission error, while the MAXENT model generated the lowest variance in variable importance. We suggest that researchers should not rely on any single algorithm, instead, they should test performance of all available models for their species and area of interest, and choose the best one to build a species distribution model.

Keywords: Generalized linear model, Kappa statistic, maximum entropy model, omission and commission error, random forest model, receiver operating characteristic curve, sensitivity, specificity, support vector machines model, true skill statistic
INTRODUCTION

Species distribution models (SDMs) are tools that integrate information about species occurrence or abundance with environmental estimates of a landscape, used to predict distribution of a species across landscapes (Elith & Leathwick 2009). When applied in a geographic information system (GIS), SDMs can produce spatial predictions of occurrence likelihood at locations where information on species distribution was previously unavailable (Václavík & Meentemeyer 2009). Though various types of algorithms are used to build different SDMs (Elith et al. 2006), they share common and general approaches (Hirzel et al. 2002) such as: (i) at a specified resolution, the study area is divided into grid cells; (ii) species presence localities (and sometimes absence localities) data are used as the dependent variable; (iii) several environmental variables (e.g., temperature, precipitation, soil type, aspect, land cover type) are collected for each grid cell as predictor variables; and (iv) the suitability of each cell for the species distributions defined as a function of the environmental variables (Stanton et al. 2012). The species distribution prediction is central to applications in ecology, evolution and conservation science (Elith et al. 2006) across terrestrial, freshwater, and marine realms (Elith & Leathwick 2009). But it remains a question for researchers which model should be selected for particular organisms and habitats of interest, particularly when few samples are present for large under-sampled areas (Mi et al. 2017).

Riparian zones are broadly defined as terrestrial landscapes with characteristic vegetation associated with temporary or permanent aquatic ecosystems (Meragiaw et al. 2018). These areas are highly complex biophysical systems, and their ecological functions are maintained by strong spatio-temporal connectivity with adjacent riverine and upland systems (Décamps et al. 2009). It has been observed that species distribution models are used more often for terrestrial environments than for aquatic or riparian ecosystems. Globally, odonates are used as model organisms to study climate change, data simulation, environmental assessment and management, effects of urbanization, landscape planning, habitat monitoring and evaluation, and conservation of rare species (Bried & Samways 2015). To date, no work has been done on the comparative use of species distribution models in India using insects as model organisms in riparian or freshwater ecosystems. With this background, in the present work we evaluated the effectiveness of four species distribution models using odonates from the Gangetic riparian zone as model organisms.

MATERIALS AND METHODS

Study area and field data collection

For the study, we selected Anax guttatus (Burmeister, 1839) commonly called Blue-tailed Green Darner (Image 1) as the model insect species. It is a dragonfly (suborder Anisoptera Selys, 1854) under the family Aeshnidae Leach, 1815 and superfamily Aeshnoidea Leach, 1815 (Dijkstra et al. 2013). The species can be identified in the field due to its large size, highly active behaviour, green colour of the thorax & first, second, & third abdominal segments, and presence of turquoise blue colour on the dorsal part of the second abdominal segment (Subramanian 2005).

We conducted the study during May 2019 from Bijnor, Uttar Pradesh to Kanpur, Uttar Pradesh (Fig. 1). The river flows through alluvial plain and covers a length of about 450km in this stretch. For the study we selected four sites, and the distance between each two successive sites was about 150km. In each site we chose a 10km river stretch and observed the presence of Blue-tailed Green Darner. We collected a total of 10 sighting locations.

Data processing and analysis

We derived the thematic layer of LULC (N.R.S.C. 2016) from multi-temporal advanced wide field sensor (AWiFS) images with 56m spatial resolution using digital and rule-based image classification methods, and forest
cover type (F.S.I. 2009) from IRS P6 (Linear Imaging Self Scanning Sensor) LISS III with 23.5m spatial resolution using a combined method of digital and on-screen visual image classification and bioclimatic layers from worldclim gridded climatic data (Fick & Hijmans 2017) with 1km spatial resolution. For analysis, we took 2km buffer zones from the river bank and resampled all the layers to 1km spatial resolution.

We used ‘stack’ function of package ‘raster’ (Hijmans 2019) to stack all the 19 available bioclimatic variable, forest cover and land use land cover (LULC) layers. After that we used ‘pairs’ function of the package ‘raster’ (Hijmans 2019) to find the correlation coefficient between stacked layers. Then we selected the variables which had a correlation coefficient less than 0.60 (Pozzobom et al. 2020), and again stacked the selected layers with ‘stack’ function of package ‘raster’ (Hijmans 2019). These selected layers were LULC, forest cover and five bioclimatic layers: annual mean temperature (Bio 1), isothermality (Bio 3), temperature seasonality (Bio 4), mean temperature of driest quarter (Bio 9), and precipitation seasonality (Bio 15).

We built four species distribution models: generalized linear model (GLM), maximum entropy (MAXENT) model, random forest (RF) model, and support vector machines (SVM).

GLM is an extension of classic linear regression modeling, where the iterative weighted linear regression technique is used to estimate maximum-likelihood of the parameters, with observations distributed in terms of an exponential family and systematic effects made linear by the suitable transformation that allow for analysis of non-linear effects among variables and non-normal distributions of the independent variables (McCullagh & Nelder 1989; Chefaoui & Lobo 2008; Shabani et al. 2016).

RF modeling is a machine learning technique which is a bootstrap-based classification and regression trees method (Cutler et al. 2007). It is used to model species distributions from both the abundance and the presence-absence data (Howard et al. 2014). It is insensitive to data distribution (Hill et al. 2017) and also takes a large number of potentially collinear variables; it is robust to over-fitting which makes it very useful for prediction (Prasad et al. 2006; Segal 2004).

MAXENT modeling is a general-purpose machine learning method to estimate a target probability distribution by finding the probability distribution of
Evaluating performance of four species distribution models

De et al.


Evaluating performance of four species distribution models

Specificity – 1 (Allouche et al. 2006) and ranges from −1 (<0.4) (Duan et al. 2014; Monserud & Leemans 1992).

Kappa statistic are as follows: excellent (0.85–1.0), very good (0.8–0.9), good (0.7–0.8), fair (0.6–0.7), and poor (0.5–0.6) (Swets 1988; Duan et al. 2014).

The area under the receiver operating characteristic curve (AUC), Kohen’s Kappa, true skill statistic (TSS), model sensitivity, model specificity, and omission error.

The true skill statistic (TSS) is expressed as Sensitivity + Specificity – 1 (Allouche et al. 2006) and ranges from −1 to +1, where +1 indicates a perfectly performing model with no error, 0 indicates the model with totally random error and -1 indicates the model with total error (Marcot 2012; Ruete & Leynaud 2015).

The model sensitivity denotes the proportion of correctly predicted presences, thus quantifying omission errors (Ward 2007; Shabani et al. 2016) and model specificity denotes the proportion of correctly predicted presences, thus quantifying commission errors (Shabani et al. 2016).

Omission error (1- sensitivity) is the under-prediction or false-negative result in areas being classified as unsuitable when they are not and commission error (1-specificity) is the over-prediction or false-positive result in areas being classified as suitable when they are not (Ward 2007) and for a good SDM, both of the omission error and commission error should be low.

For evaluation of model performance and variable importance we used ‘knitr::kable(Modelname@ evaluation’ function and ‘knitr::kable(Modelname@ variable.importance)’ function of the package ‘SSDM’ (Schmitt et al. 2017), respectively.

We chose five probability classes (0 to <0.20, 0.20 to <0.40, 0.40 to <0.60, 0.60 to <0.80 and 0.80 to 1.00) to know what percentage of the area is being declared the best and worst by each of the models by ‘ratify’ function of package ‘raster’ (Hijmans 2019).

We performed all the analysis in the ArcMap 10.3.1, QGIS 2.14.7 and in R language and environment for statistical computing (R Core Team 2019).

RESULT

The plot for each of the four models is given in Fig. 2. We found that the AUC value was highest for GLM (0.983), followed by RF (0.833), MAXENT (0.829) and SVM (0.667); the value of Kappa statistic was highest for RF (0.667), followed by GLM (0.356), SVM (0.333) and MAXENT (0.049); the value of TSS was highest for GLM (0.965), followed by RF (0.666), MAXENT (0.658) and SVM (0.334); the value of model sensitivity was 1 for GLM, 0.833 for both MAXENT and RF and 0.667 for SVM; the value of model specificity was maximum for GLM (0.965), followed by RF (0.833), MAXENT (0.825) and SVM (0.667); the omission error was lowest for GLM (0.00), for both MAXENT and RF models it was 0.167 and for SVM it was 0.333; the commission error was lowest for GLM (0.035), followed by RF model (0.167), MAXENT (0.175) and SVM (0.333) (Table 1, Fig. 3).

For GLM, RF, and SVM models the forest had...
the highest importance but for MAXENT model the Precipitation seasonality (Bio 15) had the highest importance (Table 2, Fig. 4). For GLM and SVM models the Precipitation seasonality (Bio 15) had lowest importance, for MAXENT forest had lowest importance, while for RF model Isothermality (Bio 3) had lowest importance (Table 2, Fig. 4). Overall, the variation in the variable importance was lowest in MAXENT model (SD = 3.367), followed by GLM (SD = 24.344), RF (SD = 30.868) and SVM (SD = 37.071) (Fig. 5).

By comparative analysis, we found that GLM showed 1.62% of total area as the best (occurrence probability, 0.80 to 1) and 65.50% of total area as the worst (occurrence probability, 0 to 0.20) for suitable habitat. MAXENT model showed 10.08% of total area as the best and 77.70% of total area as the worst for suitable habitat. RF model showed 5.39% of total area as the best and 23.79% of total area as the worst for suitable habitat. SVM model showed 4.53% of total area as the best and 27.68% of total area as the worst for suitable habitat (Table 3, Fig. 6).

Table 1. Values of AUC, Kappa statistic, TSS, sensitivity, specificity, omission error, and commission error generated by generalized linear model (GLM), maximum entropy (MAXENT) model, random forest (RF) model, and support vector machines (SVM) model.

<table>
<thead>
<tr>
<th></th>
<th>GLM</th>
<th>MAXENT</th>
<th>RF</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>0.983</td>
<td>0.829</td>
<td>0.833</td>
<td>0.667</td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>0.356</td>
<td>0.049</td>
<td>0.667</td>
<td>0.333</td>
</tr>
<tr>
<td>True skill statistic</td>
<td>0.965</td>
<td>0.658</td>
<td>0.666</td>
<td>0.334</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>1.000</td>
<td>0.833</td>
<td>0.833</td>
<td>0.667</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.965</td>
<td>0.825</td>
<td>0.833</td>
<td>0.667</td>
</tr>
<tr>
<td>Omission error</td>
<td>0.050</td>
<td>0.167</td>
<td>0.167</td>
<td>0.333</td>
</tr>
<tr>
<td>Commission error</td>
<td>0.035</td>
<td>0.175</td>
<td>0.167</td>
<td>0.333</td>
</tr>
</tbody>
</table>
DISCUSSION

Freshwater ecosystems, which include rivers, lakes, peat lands, swamps, fens, and springs, are highly dynamic and host a great diversity of life forms, particularly freshwater endemic species (He et al. 2019; Tickner et al. 2020). They are among the most threatened ecosystems (He et al. 2019), as globally wetlands are vanishing more rapidly than forests and freshwater species are declining faster than terrestrial or marine populations (Tickner et al. 2020). Therefore, for proper conservation management, we should understand the distribution of plants and animals inhabiting aquatic ecosystems. Species distribution models can play an important role on such efforts, because they can produce credible, defensible and repeatable information and provide tools for mapping habitats to inform decisions (Sofaer et al. 2019). Species distribution models can forecast the potential impacts of future environmental changes (Howard et al. 2014) and predict how species will respond (Buckley et al. 2010). Yet debate remains over the most robust species distribution modelling approaches for making projections (Howard et al. 2014), because these models have sensitivity to data inputs and methodological choices. This makes it important to assess the reliability and utility of the model predictions (Sofaer et al. 2019).

In the present study we compared the GLM, MAXENT, RF, and SVM approaches. We found that GLM generated the highest values for AUC, TSS, specificity and sensitivity, and the lowest values for omission error and commission error. The value of Kappa statistic was highest for RF modelling. The MAXENT model used roughly all variables equally, which is not true of the other models which put more emphasis on forest cover.

The success of a model depends on many factors, such as sample size, spatial extent of the study area, and number of ecological and statistical significant variables which affect the distribution of species of interest. We acknowledge that there were some limitations to the current work, such as that our sample size was small (only 10 presence locations), we used only seven variables, we tested only four species distribution models, and we selected a species whose distribution depends on other factors, such as the physiochemical parameters of water and availability of resources. We did not include such
Collins & McIntyre (2015) reviewed 30 studies on species distribution modelling of odonates across the world, and found that 43% used GLM, 33% MAXENT and 20% RF models. Other models used were BIOMOD, general additive model (GAM), generalized boosted model (GBM), artificial neural networks (ANN), multivariate adaptive regression splines (MARS), classified tree analysis (CTA), flexible discriminant analysis (FDA), boosted regression trees (BRT), surface range envelopes (SRE), and mixture discriminant analysis (MDA). Different species distribution models produce different results (Shabani et al. 2016), and the same model can give different results for different species and areas. We urge researchers not to rely on just one model, rather they should compare different available species distribution models and select the best one.

Our study was in India where an insect was used for comparative evaluation of species distribution models in a riverine riparian zone. We recommend that further
studies on different species distribution models using different animals and ecological variables should be carried out in the riparian zones of Indian river systems for proper design and implementation of ecological habitat management plans.

REFERENCES


Evaluating performance of four species distribution models

De et al.


The total number of unique words is 5277.
Elevational pattern and seasonality of avian diversity in Kaligandaki River Basin, central Himalaya
– Juna Neupane, Laxman Khanal, Basant Gyawali & Mukesh Kumar Chalise, Pp. 16927–16943

Species diversity and feeding guilds of birds in Malaysian agarwood plantations
– Nor Nasibah Mohd Jamil, Husni Ibrahim, Haniza Hanim Mohd Zain & Nur Hidayat Che Musa, Pp. 16954–16961

Evaluating performance of four species distribution models using Blue-tailed Green Damner Anax guttatus (Insecta: Odonata) as model organism from the Gangetic riparian zone

Butterfly species richness and diversity in rural and urban areas of Sirajganj, Bangladesh

Chroococcalean blue green algae from the paddy fields of Satara District, Maharashtra, India
– Sharada Jagannath Ghadage & Vaneeta Chandrashekhar Karande, Pp. 16979–16992

A rare camera trap record of the Hispid Hare Caprolagus hispidus from Dudhwa Tiger Reserve, Terai Arc Landscape, India
– Sankarshan Rastogi, Ram Kumar Raj & Bridesh Kumar Chauhan, Pp. 17024–17027

First distributional record of the Lesser Adjutant Leptoptilos javanicus Horsfield, 1821 (Ciconiiformes: Ciconiidae) from Sindhuli District, Nepal

First record of African Sailfin Flying Fish Parexocoetus mento (Valenciennes, 1847) (Beloniformes: Exocoetidae), from the waters off Andaman Islands, India

A first distribution record of the Indian Peacock Softshell Turtle Nilssonia hurum (Gray, 1830) (Reptilia: Testudines: Trionychidae) from Mizoram, India
– S. Dam Roy, Pp. 17032–17035

A frog that eats foam: predation on the nest of Polypedates sp. (Rhacophoridae) by Euphlyctis sp. (Dicroglossidae)

New distribution record of two endemic plant species, Euphorbia kadapensis Sarojin. & R.V.R. Raju (Euphorbiaceae) and Lepidagathis keralensis Madhus. & N.P. Singh (Acanthaceae), for Karnataka, India

A rare camera trap record of the Hispid Hare Caprolagus hispidus from Dudhwa Tiger Reserve, Terai Arc Landscape, India
– Sankarshan Rastogi, Ram Kumar Raj & Bridesh Kumar Chauhan, Pp. 17024–17027

First distributional record of the Lesser Adjutant Leptoptilos javanicus Horsfield, 1821 (Ciconiiformes: Ciconiidae) from Sindhuli District, Nepal

First record of African Sailfin Flying Fish Parexocoetus mento (Valenciennes, 1847) (Beloniformes: Exocoetidae), from the waters off Andaman Islands, India

A first distribution record of the Indian Peacock Softshell Turtle Nilssonia hurum (Gray, 1830) (Reptilia: Testudines: Trionychidae) from Mizoram, India
– S. Dam Roy, Pp. 17032–17035

A frog that eats foam: predation on the nest of Polypedates sp. (Rhacophoridae) by Euphlyctis sp. (Dicroglossidae)

New distribution record of two endemic plant species, Euphorbia kadapensis Sarojin. & R.V.R. Raju (Euphorbiaceae) and Lepidagathis keralensis Madhus. & N.P. Singh (Acanthaceae), for Karnataka, India

Cirsium wallichii DC. (Asteraceae): a key nectar source of butterflies
– Bitupan Boruah, Amit Kumar & Abhijit Das, Pp. 17049–17056

Hypecoum pendulum L. (Papaveraceae: Ranunculales): a new record for the flora of Haryana, India

A rare camera trap record of the Hispid Hare Caprolagus hispidus from Dudhwa Tiger Reserve, Terai Arc Landscape, India
– Sankarshan Rastogi, Ram Kumar Raj & Bridesh Kumar Chauhan, Pp. 17024–17027

First distributional record of the Lesser Adjutant Leptoptilos javanicus Horsfield, 1821 (Ciconiiformes: Ciconiidae) from Sindhuli District, Nepal

First record of African Sailfin Flying Fish Parexocoetus mento (Valenciennes, 1847) (Beloniformes: Exocoetidae), from the waters off Andaman Islands, India

A first distribution record of the Indian Peacock Softshell Turtle Nilssonia hurum (Gray, 1830) (Reptilia: Testudines: Trionychidae) from Mizoram, India
– S. Dam Roy, Pp. 17032–17035

A frog that eats foam: predation on the nest of Polypedates sp. (Rhacophoridae) by Euphlyctis sp. (Dicroglossidae)

New distribution record of two endemic plant species, Euphorbia kadapensis Sarojin. & R.V.R. Raju (Euphorbiaceae) and Lepidagathis keralensis Madhus. & N.P. Singh (Acanthaceae), for Karnataka, India

Cirsium wallichii DC. (Asteraceae): a key nectar source of butterflies
– Bitupan Boruah, Amit Kumar & Abhijit Das, Pp. 17049–17056

Hypecoum pendulum L. (Papaveraceae: Ranunculales): a new record for the flora of Haryana, India

Addendum

Erratum and addenda to the article ‘A history of primatology in India’