A Machine Learning Based Approach to Predict Species-Habitat Relationship in Australia

MSc Research Project
Data Analytics

Preethi Belur Ramesh
Student ID: x20180101

School of Computing
National College of Ireland

Supervisor: Bharat Agarwal

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Preethi Belur Ramesh
Student ID:	x20180101
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Bharat Agarwal
Submission Due Date:	15/12/2022
Project Title:	A Machine Learning Based Approach to Predict Species-Habitat Relationship in Australia
Word Count:	4734
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	1st February 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Machine Learning Based Approach to Predict Species-Habitat Relationship in Australia

Preethi Belur Ramesh
x20180101

Abstract

The ecological balance hangs on the abundance of flora and fauna. The declining numbers of species can be attributed to multiple reasons such as climate change, habitat destruction and human interference. Therefore, this research aims to represent the relationship between the species and the climatic conditions of the habitat using Machine Learning Techniques and Statistical Analysis to understand the climatic conditions affecting the distribution patterns of the endemic species of Australia. As a novel approach, this research also tries to draw parallels between the warm-blooded bird species and cold blooded frog species to infer if species that inherently react differently to the climate show similar signs of decline and distribution to historic and future climate variables.

1 Introduction

An understanding of the species-habitat relationship is necessary for efficient conservation and management decisions. Modern ecological theories generalize ecological patterns over geographical regions and attribute spatial distribution of species to climatic conditions, vegetation, and food availability (Bagariaa et al.; 2021). The decreasing distribution or abundance of species caused by climate change has disrupted the species-habitat associations, according to most conservationists and wildlife organizations. Furthermore, rising seasonal temperatures that correlate with poor habitat quality are responsible for the phenological anomalies in species Foden et al. (2019). About 66% of ectothermic species and 83% of endothermic species are highly susceptible to climate change impacts which are not identified by the the International Union for Conservation of Nature (IUCN) Red List of Threatened Species indicating that the species are under a larger threat than the conservation status might alone indicate (Hart et al.; 2018). According to (Hoffmann et al.; 2019), Australia is prone to notable climate-change factors such as bushfires, extreme temperature and rainfall and droughts. Due to these factors, forest and woodland areas are likely to be reduced, heatwaves occur, and snow cover in the Australian Alps is reduced, putting both marine and terrestrial ecosystems at risk. Furthermore, the distributions of species belonging to different taxonomic groups might change differently due to environmental factors (Hart et al.; 2018). It is especially true for endotherms and ectotherms because their phenotypes and interactions with the environment differ.

This study explores one of the most powerful mechanisms, Species Distribution Models, that can be used to investigate the correlation between climate variables and species

Table 1: Ecological Terms Used in the Research

Terms	Definition
Endotherms	Warm-blooded species that can regulate the internal temperature using body heat
Ectotherms	Cold-blooded species that rely on external sources to control the body temperature
Endemic Species	Native to a particular country or region
Genus	Biological classification that groups similar species.
Phenotype	Characteristics of an organism based on its interaction with environment

spatial distributions. The approach is based on the use of robust machine learning algorithms coupled with statistical models on species occurrence data associated with climate data to understand changes in species' habitats as a result of environmental drivers. Moreover spatial projections of changes in habitat and distribution as a result of present and future climate change impacts are analyzed in depth.

In particular, this research examines the distribution of Australian endemic species, based on historic bio-climatic variables, by using machine learning techniques such as Boosted Regression Trees and Random Forests to determine the presence and absence of these species in Australian geographical spaces. Also, the response curves are analyzed to determine the climatic variable that is more predominantly responsible for the change in species habitat. Further, the Ensemble model is utilized to forecast the distribution of species in the future based on the future climate data with carbon emission concentrations projected at Radiative Concentration Pathways (RCP) 8.5 ppm, a value derived from multiple climate models. Comparing the habitat behavior of ectothermic frog and endothermic bird species in relation to climate change is another significant part of this research. Table 1 represents the ecological terms used in this research as a helpful guide.

The research paper is divided into following sections, firstly, the existing research on the species distribution models are critically analyzed, the next section represents the general design framework that the species distribution models follow along with the framework followed in this research. The fourth section delves deep into the methodology and implementation of the models followed by the model evaluation. Finally, the discussion and future work describe the interpretations of the findings along with the research implications.

2 Related Work

This section critically analyzes the previous work in the prediction of species distribution based on climate variables. The contributions and limitations of each paper is described in depth. Moreover, the modelling methods, evaluation techniques and Machine Learning and Statistical Approaches undertaken are explored abundantly.

2.1 Species Distribution and Climate Impacts

The article (Hart et al.; 2018) analyses the reduced population of the Tufted Puffin species and predicts the habitat shifts due to warming marine and land temperatures using the species distribution modelling. Past and Future climatic data are scaled and used as explanatory variables against the species data to model the species' spatial distribution changes by including generalized boosting models, random forests, generalized linear models, and ensemble modelling techniques. Although the study depicts the effects of warming temperatures on population trends, it is limited to only a single species. It does not provide insights into the habitat suitability differences amongst other related seabird species. Furthermore, the SDM implementation does not include any replication techniques such as cross-validation or bootstrapping that potentially reduce the model uncertainty.

The study (Bagariaa et al.; 2021) includes 21 species occurring in the Himalayan arc. The species are grouped into different cohorts by performing ordination and clustering techniques upon which 2 different modelling frameworks such as SDM and Bioclimatic Envelope Models (BEM) were applied. The results were used to access the clusters prone to decline due to climate change. The only limitation is that there are uncertain ecological relationships with clusters having rare species owing to the lesser occurrences.

Another work (Shabani et al.; 2018) analyzes the accuracy of species distribution models generated using five types of bio-climatic models, and three threshold selections to predict the distribution of eight plant species in Australia. The study limits to the discussion of the accuracy of models without indulging in what those results mean in terms of the habitat suitability for the plant species, however, the study indicates TSS to be a better evaluator for different models compared to AUC.

In (Tiago et al.; 2017) species distribution modelling is performed on 20 species consisting of 11 reptiles and 9 amphibians. The modelling technique used is Generalized Additive Models to get an outcome of 10 replicate predictions for data obtained from two different sources to compare if the model performance depended upon species having well-sampled climatic niches. The SDM evaluation is limited in the paper but gives a good overview of factors that affect SDM performance.

(Dyderski et al.; 2018) examine the effects of climatic changes on the European forest tree species. The implementation is carried out by developing General Circulation models for three climate change scenarios such as optimistic, moderate and pessimistic. The model outcomes were evaluated using AUC and it was observed that the plant species could be categorised into a winner, loser and alien groups based on the responses of the tree species to climate change. This study intricately presents the feature importance of the 3 categories but fails to look at threshold-dependent evaluation techniques and the prevalence of species.

2.2 Analysis of Model Performance

The increasing uncertainty of species distribution models due to collinearity issues amongst the predictor variables are discussed in (De Marco and Nobrega; 2018). The SDM models are implemented using both collinear predictors and non-collinear predictors derived from PCA, the predictions are derived from both models to compare the spatial distribution of species. The outcome of the study concludes that the PCA-derived variables control the negative impacts of collinearity on the models.

Authors of (Fourcade et al.; 2018) explore the performance of species distribution modelling using pseudo-environmental predictors instead of real bio-climatic variables. The pseudo-predictors were created using the classical paintings obtained from Google search. The results show that the SDMs could easily be derailed using meaningless predictor variables and show that the evaluation of SDMs cannot rely only on AUC. It also indicates the importance of selecting only those environmental predictors that can add value to the models.

Research by (SUNG et al.; 2018) addresses the effectiveness of the models when species of different sample sizes are included. A total of seven sampling sizes and eight modelling techniques are used along with ANOVA, a statistical technique used to differentiate the models. It was noted that the model performance plateaued as the sample size became greater than 200. This study informs the correct use of statistical techniques in SDMs, then again relies only on AUC for evaluating the model performance.

3 Design Framework

The general design of Species Distribution Modelling framework includes components such as (1) The species occurrence data, (2) location representation of the environmental characteristics, and (3) a model that maps the species observations and environmental characteristics to predictions (Beery et al.; 2021).

Formally the components are defined as follows,

(1) The species observation or occurrence data represents a collection of records indicating the presence or absence of species at a particular location. This can be written as $\{x^i, y^i\}_i^N$ where $x^i \in$ S or spatial location represented using latitudes and longitudes, and $y^i \in \{0, 1\}$ where absence(0) and presence(1).

(2) A location representation of environmental characteristics $h(x) \in R^k$ where k is geospatial data layers or climatic layers and $h(x)$ is a way of representing the species observations $x \in S$ in this environmental location. In other words, inclusion of the environmental feature space (Australian climatic layers) where the species (endemic birds & frogs) are observed.

(3) Model $\int_\theta : Z \longrightarrow [0, 1]$ where θ is a parameter vector. The objective of the aforementioned function $\int_\theta(h(x))$ is to predict the presence and absence of species on a geospatial location using a supervised learning machine learning model. In other words, predicting distribution of endemic species by mapping interpolated climatic layers in Australia using ML models.

The methodology followed in this research is represented in the form of a flowchart in Figure 1. The approach is along the lines of the formal definition described in the above section. Each step of the methodology such as the Spatial Data Extraction, Spatial Data Pre-processing, Model Generation, Model Resampling, Model Evaluation, and Result Interpretation will be explained in detail in the next sections. Finally, the similarities and differences in the distribution of endothermic and ectothermic species will be compared.

Further, forecast of the species distribution is predicted using the weighted ensemble techniques with statistical methods such as Test Skill Statistic(TSS) and AUC that takes the future climatic geospatial data combined with the model objects generated for each species group as input (depicted in Figure 2).

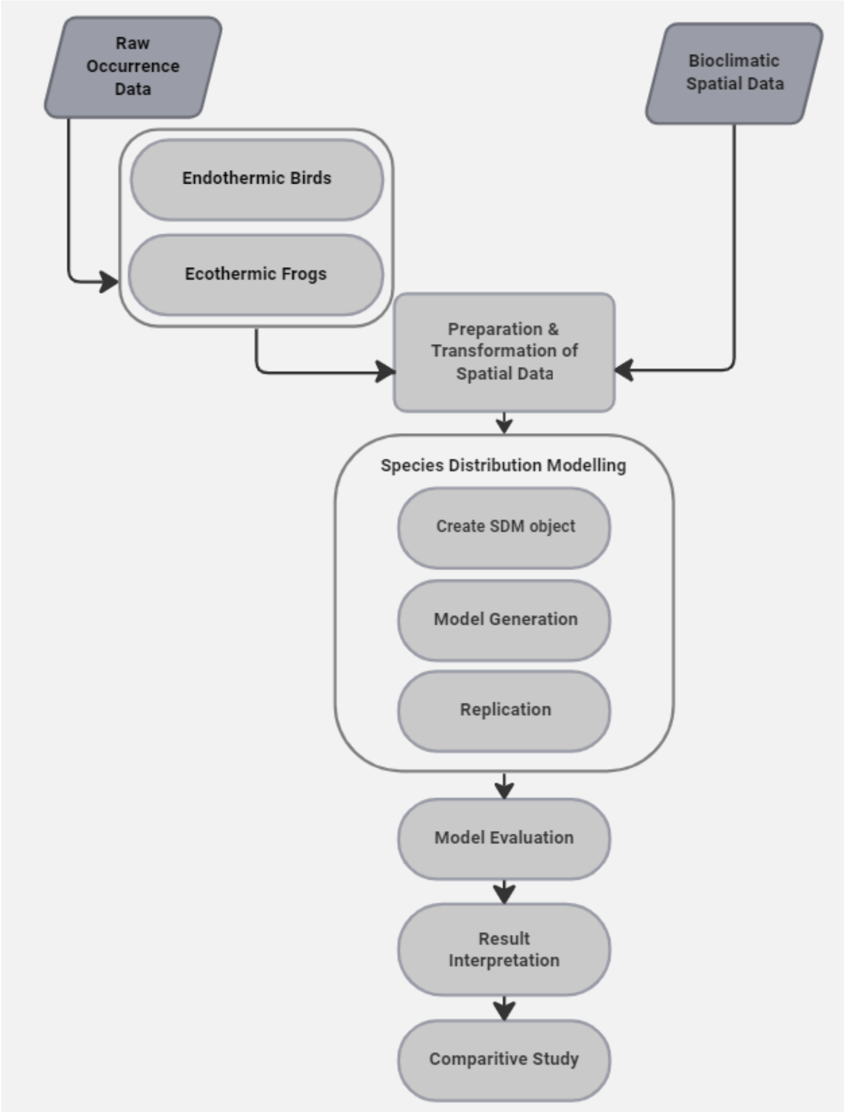


Figure 1: Flowchart: Design Methodology of Endemic Species Distribution

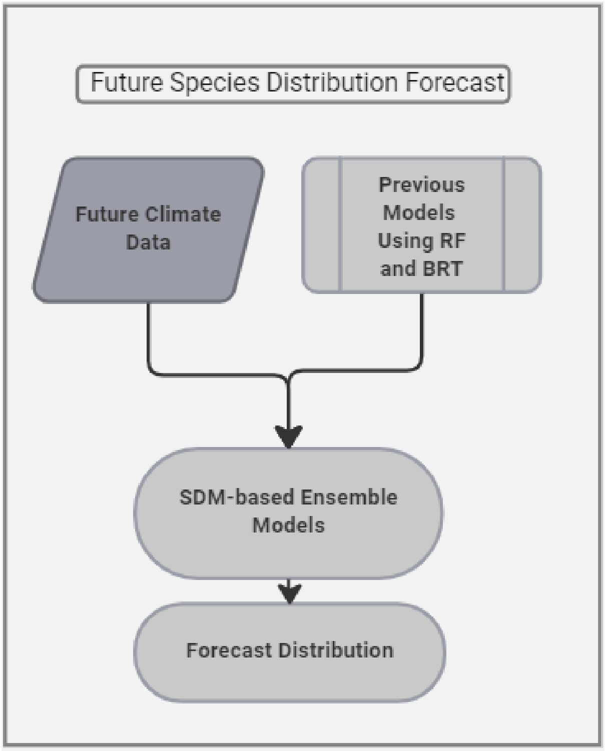

Future Species Distribution Forecast
Future Climate Data
Previous Models Using RF and BRT
SDM-based Ensemble Models
Forecast Distribution

Figure 2: Flowchart: Forecasting Endemic Species Distribution in the Future

4 Methdology

4.1 Extraction of Species and Environmental Data

4.1.1 Historical Climate Data

The Environmental data consisting of the bioclimatic variables or the biologically meaningful variables defined for the period 1970-2000 are obtained from the WorldClim[1] at a spatial resolution, 5 arc minutes of a latitude/longitude degree which is about 9km at the equator using the *getData* method, *Raster*[2] package in R. The data obtained is WorldClim version 2.1 released in January 2020. The 19 bio-climatic variables (described in Table 2) consist of a set of global climate layers resulting from interpolation of the monthly climate observations.

4.1.2 Endemic Species Data

The Australian endothermic bird species and the ectothermic frog species occurrence data is obtained from the Global Biodiversity Information Facility (GBIF)[3] combining datasets from resources such as EOD – eBird Observation Dataset, NSW BioNet Atlas, BirdLife Australia, Birdata, Victorian Biodiversity Atlas, SA Fauna (BDBSA), WildNet - Queensland Wildlife Data, Gang-gang cockatoo survey, Fauna Atlas N.T, iNaturalist Research-grade Observations, New South Wales Bird Atlassers, FrogID, Australian Museum provider for OZCAM, and South Australian Museum. Furthermore, the species listed by the IUCN as Critically Endangered, Endangered, Vulnerable and Near Threatened

[1] www.WorldClim.org

[2] https://www.rdocumentation.org/packages/raster/versions/3.6-11

[3] www.gbif.org

Table 2: Bio-Climatic Variables Obtained From WorldClim[1]

Bio-Climatic Variables	**Description**	**Unit**
BIO1	Annual Mean Temperature	°C
BIO2	Mean Diurnal Range	°C
BIO3	Isothermality (BIO2/BIO7) (* 100)	°C
BIO4	Temperature Seasonality (standard deviation *100)	°C/100
BIO5	Warmest Month's Maximum Temperature	°C
BIO6	Coldest Month's Minimum Temperature	°C
BIO7	Annual Range of Temperature (BIO5-BIO6)	°C
BIO8	Wettest Quarter's Mean Temperature	°C
BIO9	Driest Quarter's Mean Temperature	°C
BIO10	Warmest Quarter's Mean Temperature	°C
BIO11	Coldest Quarter's Mean Temperature	°C
BIO12	Annual Precipitation	kg m-2
BIO13	Wettest Month's Precipitation	kg m-2
BIO14	Driest Month's Precipitation	kg m-2
BIO15	Precipitation Seasonality (Coefficient of Variation)	kg m-2
BIO16	Precipitation in the Wettest Quarter	kg m-2
BIO17	Precipitation in the Driest Quarter	kg m-2
BIO18	Precipitation in the Warmest Quarter	kg m-2
BIO19	Precipitation in the Coldest Quarter	kg m-2

only are included in this study. Overall, the occurrence data has 257 columns with crucial information such as the count of the individual species and coordinates of the species' presence.

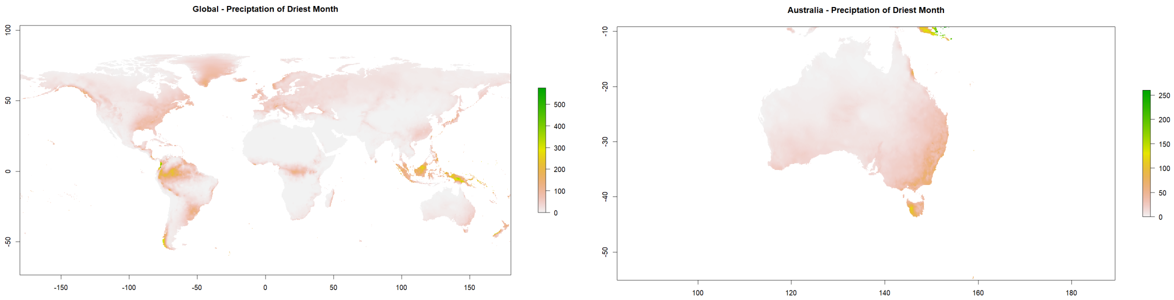
4.1.3 Future Climate Data

The Future climate data is obtained similar to the historical climatic data using the *getData* method from the Coupled Model Intercomparison Project 5 (CMIP5). Additional arguments must be included in the method such as the resolution of the climate data, Representative Concentration Pathways (RCPs) which represents the concentration of carbon dioxide and other greenhouse gases measured in parts per million, time period in years and model using which the climate data is interpolated. The parameters that were chosen for this study are 5 arc minute resolution, RCP of 8.5, 70 years and multi-ensemble model.

4.2 Preparing And Transforming Spatial Data

The Data Preparation step involves gathering and processing the biodiversity and environmental data.

For the best SDM results, spatial coordinates that indicate the presence or existence of species and the absence of species must be consistent between the environmental and species data. Furthermore, it is imperative to have occurrence data containing both presences and absences or pseudo-absences or background points having proper sample sizes (Liu et al.; 2018). Therefore, the endothermic and ectothermic species' occurrence data records with null latitude or longitude values or individual species counts less than



(a) Global Climatic Regions

(b) Australian Climatic Regions

Figure 3: Global Climatic Regions Cropped to Include Only Australia

5 in a particular region are not considered to reduce uncertainty in the model. However, by doing so, the absence data points were reduced. This issue is resolved while creating an SDM object.

Finally, the biodiversity data is transformed to contain only endemic (native or found only in Australia) species. After this transformation, the data consisted of 47 species of birds and 22 species of frogs.

The columns latitude, longitude, and occurenceStatus are extracted from the species data to create a new dataframe. Furthermore, the latitude and longitude columns are converted into a spatial coordinates object by *coords* method and the occurrenceStatus which represents 'presence' or 'absence' of species is converted into a factor of 0s and 1s. Due to the former conversion into a spatial object, the dataframe is automatically type-casted into a Spatial Points Dataframe.

The bio-climatic data of type *RasterStack* contain not just the Australian regions but the global climatic layers. Therefore, the global climatic regions were cropped only to include the spatial polygons that is 112.92°W to 159.1°W and -55.1°N to -9.14°N of Australia as the species are endemic in nature.

Figure 3 represents one of the bio-climatic variables, precipitation of the driest month (BIO14), cropped to include only the Australian spatial range.

Since bio-climatic variables are interpolated from climate observations, there is a tendency for them to be multicollinear. A combination of two strategies is implemented to tackle this (Naimi and Araújo; 2016). The first procedure, *vifcor* includes iteratively finding pairs of variables with maximum linear correlation or having a correlation higher than the threshold. Out of the pair, one variable with greater Variance Inflation Factor (VIF) is removed. This procedure is carried out until no multicollinear variable remains. The second strategy, *vifstep* includes a stepwise procedure of calculating the VIF of all the variables while removing the one with the highest VIF at each step until no variables are highly correlated to one another. The common variables that remain after both the procedures are considered for the modelling.

Table 3 represents the remaining bio-climatic variables after performing the aforementioned strategies and Figure 4 depicts the spatial representation of these variables.

Although multiple studies have included Principle Component Analysis (PCA) to eliminate the highly collinear variables, it was not considered in this study for two reasons (Gonzalez; 2018), (1) Including PCA reduces the interpretability of the feature importance that are derived from the modelling, and (2) the multicollinear variables reduction techniques followed by PCA complicates the projection of new geographical or temporal extent.

Table 3: VIF of The Remainder Bio-Climatic Variables

(a) vifstep

Variables	VIF
BIO2	2.295498
BIO3	4.405041
BIO8	2.461119
BIO9	3.430463
BIO13	5.195553
BIO14	3.508061
BIO15	2.254147
BIO18	4.178966
BIO19	3.260141

(b) vifcor

Variables	VIF
BIO2	2.074739
BIO8	2.240553
BIO9	2.224379
BIO13	4.89417
BIO14	3.174495
BIO15	2.157783
BIO18	3.877207
BIO19	2.955217

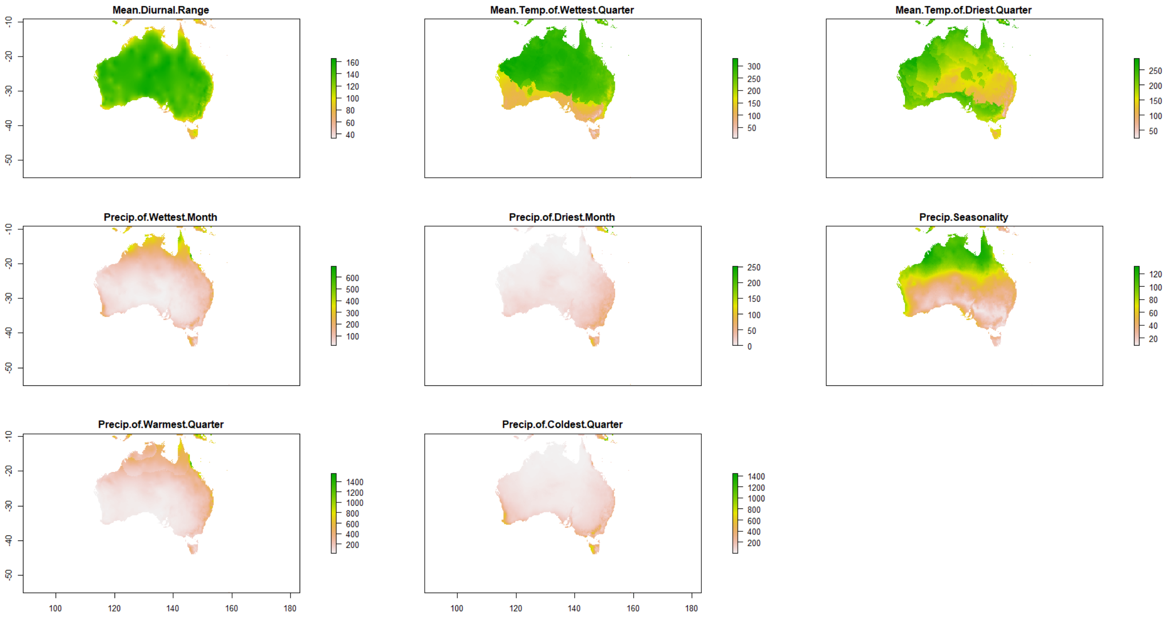
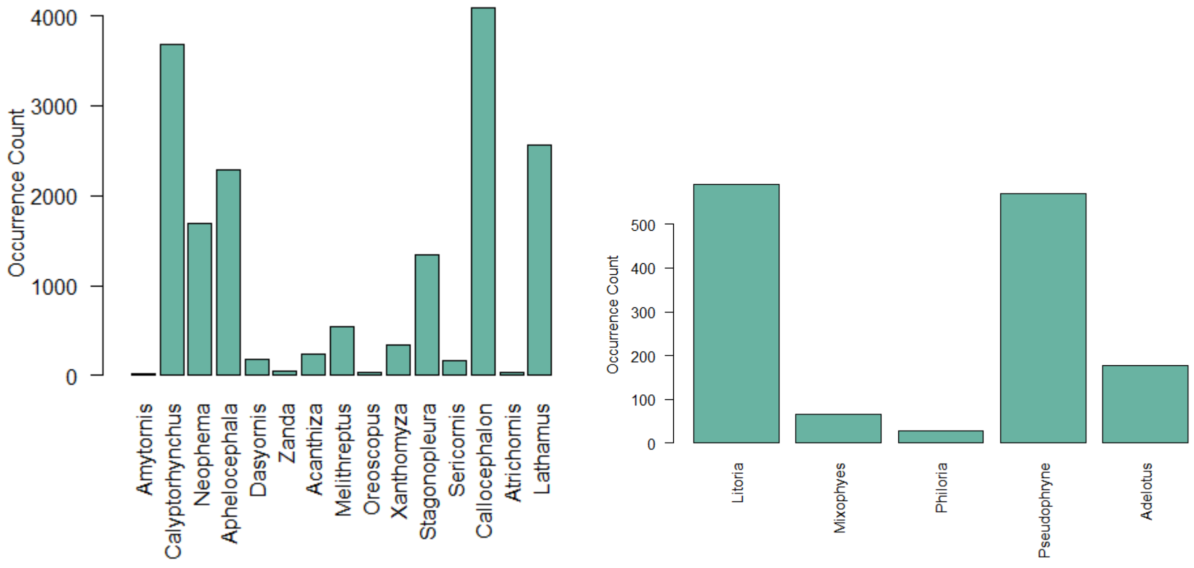


Figure 4: Bio-Climatic Variables After Removing Multicollinearity

4.3 Modelling The Distribution Of Endemic Species

The 47 species of birds are further combined into 15 different groups based on the genus. For example, *Calyptorhynchus baudinii*, *Calyptorhynchus lathami*, and *Calyptorhynchus latirostris* are grouped together as they belong to the genus *Calyptorhynchus*. Similarly 22 species of frogs were combined into 5 groups. Figure 5 shows the occurrence counts of species data included in this study.



(a) Endotherms

(b) Ectoherms

Figure 5: Occurrence Count - Endotherms vs Ectotherms

Some of the related works have used the *BIOMOD* or *dismo* packages in R. However, these packages have limitations in terms of the number of species distribution related packages and modelling methods. Furthermore, the packages lack standardized modelling techniques which makes it difficult to compare the outcomes to determine the superior technique while modelling the species distribution. In this research, the *sdm* package in R released recently by (Naimi and Araújo; 2016) was used to model the species distribution as it is a unified and standardized framework that supports several modelling approaches that are both mechanistic and correlative. Furthermore, the sdm is object-oriented in design, making it an efficient error handler in addition to being highly readable due to its graphical user interface built on *Rshiny*.

4.4 Model Generation

The *sdm* library consists of the *sdmData* method used to create a data object of single or multiple species. It includes arguments such as explanatory variables (bio-climatic variables), background data and species observation points.

One of the main concerns while analyzing the species distribution are the sparsely available absence data points. To compensate for the reduction of absence data points while removing the inconsistent data, the *gRandom* was used to create pseudo-absences or background points. The method works by assigning coordinates in a geographical space as the absence sites of the species while ensuring that these points do not match with the known present points. Also, the number of background points is decided based on the sample size of species belonging to a particular genus.

The data objects created are then fit and evaluated using the *sdm* method. This method is responsible for model building using Machine Learning Techniques with underlying statistical methods. It requires arguments like the data object, techniques to fit the models, and optional replication or resampling methods with train and test split.

The Boosted Regression Trees (BRT) and Random Forest (RF) techniques are chosen to perform the species distribution modelling. BRT, a combination of decision tree and boosting methods improves the predictive performance by building weighted trees with a random subset of input data while improving upon the weaker learners by recognizing the errors in the previously fitted to build the subsequent trees. On the other hand, the RF is built upon several trees using a bagging approach having a different combination of the original data in each tree, which leaves room for less error. Moreover, both these techniques help realize the most important climatic layers that impacts the distribution of species (Hart et al.; 2018).

4.5 Model Resampling

In order to make the model predictions less biased, it is important to include multiple representations of the samples of the population (Dodangeh et al.; 2020). While predicting the species distribution, resampling techniques such as bootstrapping, cross-validation, and sub-sampling can be used to partition the data into dependent or independent test percentages and run the model multiple times to obtain best results.

In this research, cross-validation is used as a resampling technique because it increases the predictive nature of the modelling while having climate variables that are strongly correlated. In this approach, the model is fit on the training data while the evaluation is performed on the test. As independent test data is not available, the data is partitioned into 80% training and 20% dependent test data. Furthermore, the data is divided into 5 folds where one part is used for testing and four parts are used to train the model (Valavi et al.; 2019).

5 Evaluating the Model Performance

The models generated using BRT and RF techniques with 5-fold cross validation can be evaluated based on the model output. Generally, the distribution models produce both binary and continuous results. The binary results categorize the geographical space as either containing species distributions or not, whereas the continuous results indicate the probability of such distributions occurring.

The binary results of SDM can measure the discrimination capacity of the model, which showcases the model's ability to correctly classify the presence and absence sites of the species. On the other hand, the continuous results can measure the reliability of the model which determines the conformance between the predicted probability of occurrence and proportion of observed occurrences Liu et al. (2009).

In order to evaluate either discrimination capacity or reliability, a range of indices can be used, of which there are two types - (1) Threshold-dependent indices and (2) Threshold-independent indices.

5.1 Threshold-Dependent Evaluation

The Threshold-Dependent evaluation is performed using indices that work on binary results or on continuous results that have been converted into binary outcome with the aid of a specific cut-off value, termed as threshold.

As a general rule of thumb, 0.5 or 0.05 are commonly considered to be the threshold in ecology to determine the model accuracy. However, since there is no ecological significance in choosing such thresholds, the best approach to choose a threshold is the one that maximizes the agreement between the predicted and observed distributions.

(1) True Skill Statistic (TSS) - One of the effective threshold-dependent measure of accuracy of the distribution of species as it provides equal weights to sensitivity and specificity. Furthermore, since TSS does not depend on prevalence, which is the frequency of species occurrence, TSS can evaluate models without creating bias based on sample size (Kumar et al.; 2018).

$$TSS = Sensitivity + Specificity - 1 \tag{1}$$

In general, a model's performance is considered better if the TSS value approaches 1 or is higher than threshold of 0.7. Figure 6 depicts lollipop plots indicating the mean performance of TSS on the test data for the different groups of endemic bird and frog species.

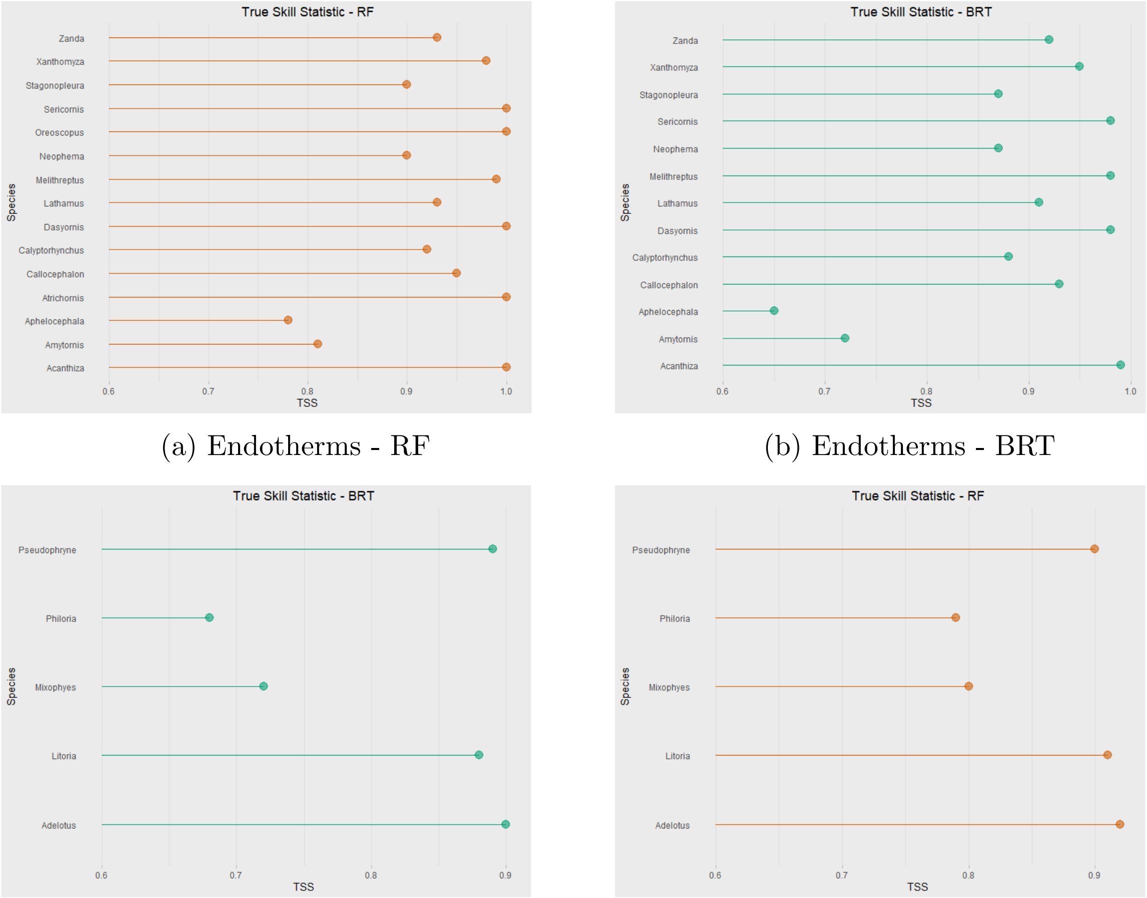
The following observations were made comparing the models and the species, It is notable that the models generated using RF performs better than the BRT in all instances for both species. Additionally, in endotherms, both models seem to perform relatively bad for genus groups *Aphelocephala* and *Amytornis* which cannot be to attributed to a valid ecological reason. Similarly, in ectotherms, the models generate lower TSS for species groups belonging to *Mixophyes* and *Philoria* which could be attributed to their sample size. Overall, the models have performed relatively well for endotherms compared to ectotherms.

5.2 Threshold-Independent Evaluation

The Threshold-Independent evaluation is performed with indices that can be applied on the continuous results without having to convert them to a binary solution.

(1) Area Under Curve (AUC) - The Area Under Curve are obtained from the ROC plot and similar to TSS, are insensitive to prevalence. The ROC plot is generated by mapping the probability of predicted occurrence against the commission error. Additionally, AUC can be calculated for all thresholds that fall under the probability of predicted occurrence, regardless of which threshold is used (Jiménez and Soberón; 2020). In other words, an AUC value indicates the likelihood that, if a presence site and absence sites are randomly selected from the population, the predictability of the presence site is higher than an absence site.

Figure 7 depict the ROC plots for the endemic bird species for the models RF and BRT. These plots are generated on all the species group at a time with the cross-validation. The AUC is measured for both training and test data indicated by colors red and blue respectively. Although both models perform well (indicating that both presence and absence sites are well-distinguished), the RF model performs better having 0.997 on training and 0.963 on test data.



(a) Endotherms - RF

(b) Endotherms - BRT

(c) Ectotherms - RF

(d) Ectotherms - BRT

Figure 6: Mean Model Performance based on True Skill Statistic - RF vs BRT

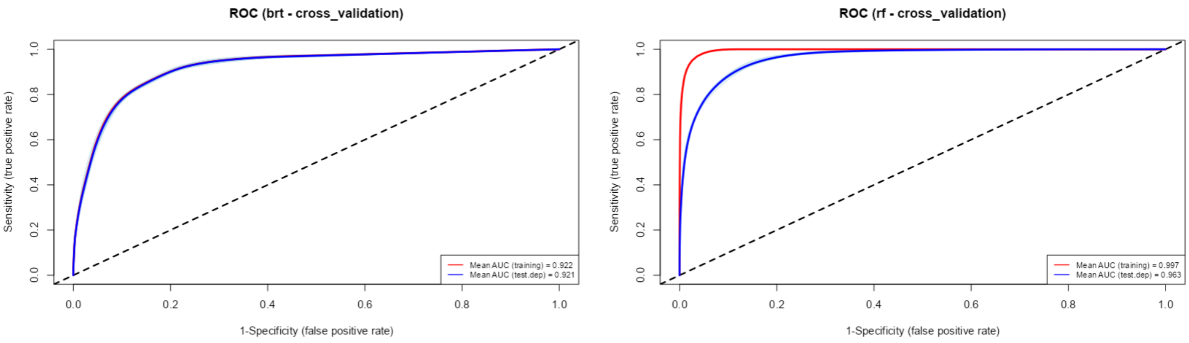


Figure 7: Mean Model Performance based on ROC plot - RF vs BRT

Figure 8 depict the ROC plots for the endemic frog species for the models RF and BRT. These plots are generated on all the species group at a time with the cross-validation. The AUC is measured for both training and test data indicated by colors red and blue respectively. Although both models perform well (indicating that both presence and absence sites are well-distinguished), the RF model performs better having 0.997 on training and 0.963 on test data.

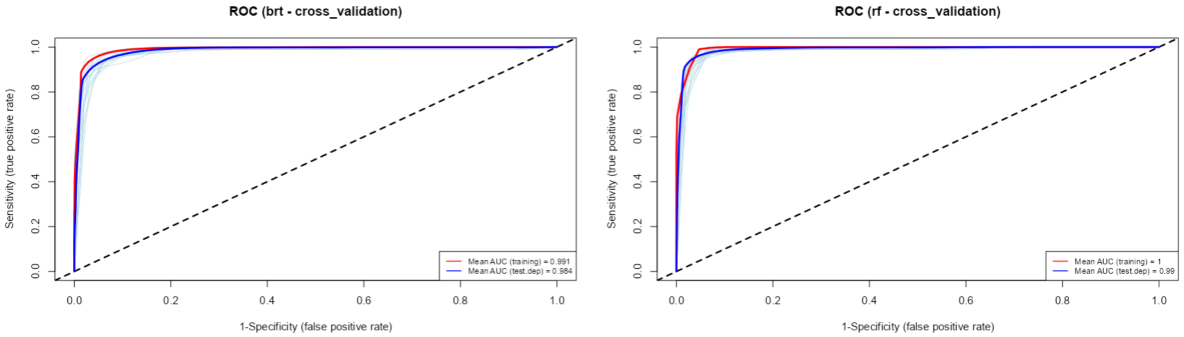


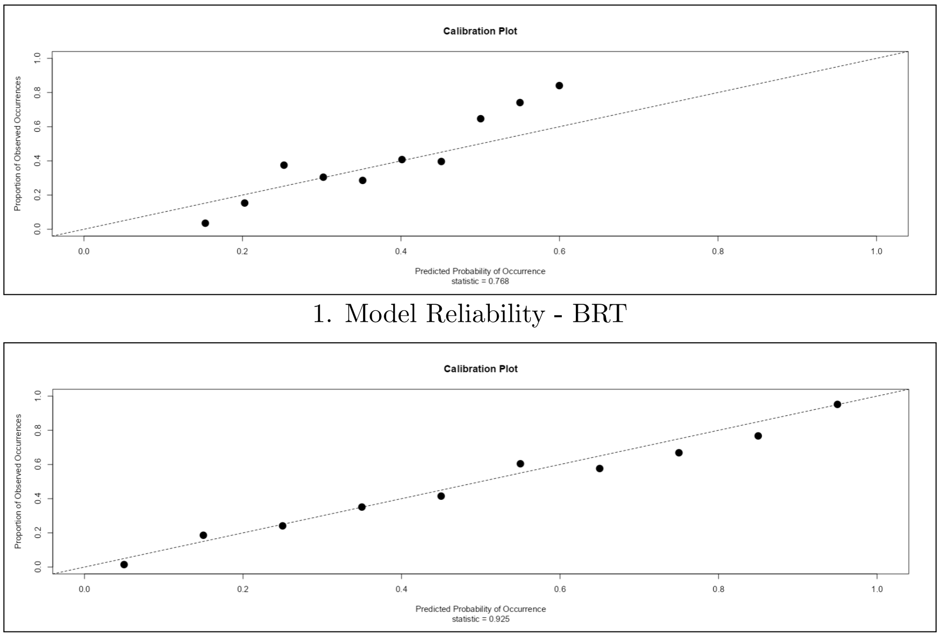
Figure 8: Mean Model Performance based on ROC plot - RF vs BRT

(2) Model Calibration - The Model Calibration plots measure the reliability of both RF and BRT models by considering only the test data. This measure is built on the distribution models represented as continuous results. Despite higher AUCs in both cases, the calibration plots (shown in Figure 9) indicate only 0.786 for BRT and 0.925 for RF of the predicted probability of occurrence. Of the two models, RF turns out to have better predictive ability.

5.3 Response Curves

The response plots indicated in Figure 10 depict the relationship that each species group have between their probability of occurrence and the bio-climatic variables. The response plots are generated for only one bio-climatic variable while having the other variables maintained at a constant. The values on the Y-axis represent the probability of occurrence of species ranging from 0.2 to 0.6 and the X-axis represent the scale for the environmental variables.

The response curves indicate that the probability of occurrence of endotherms is higher as the precipitation increases (bio19) and that the species decline with increasing tem-



1. Model Reliability - BRT

2. Model Reliability - RF

Figure 9: Calibration Plots Indicating Model Reliability

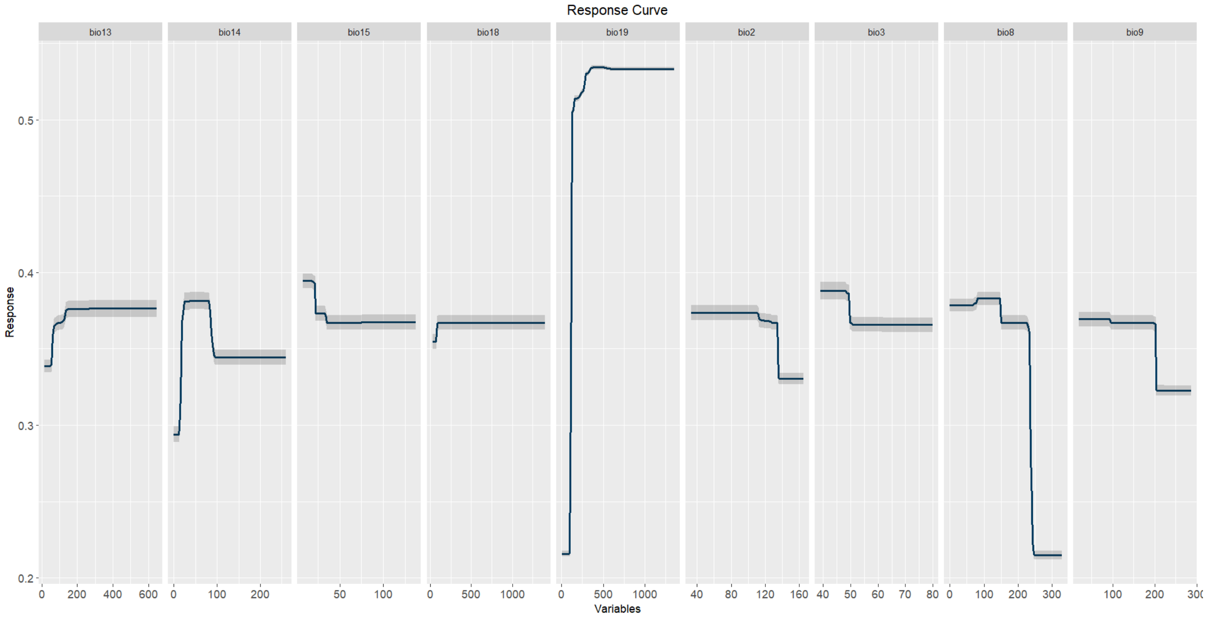


Figure 10: Response Curves - Endotherms

perature (bio8), indicating that endotherms prefer wetter areas with low temperatures.

Similarly, the response curves indicate that the probability of occurrence of ectotherms is higher as the precipitation of increases (bio14) and (bio19) and that the species decline with increasing mean temperature (bio8), indicating that ectotherms prefer wetter areas with low temperatures.

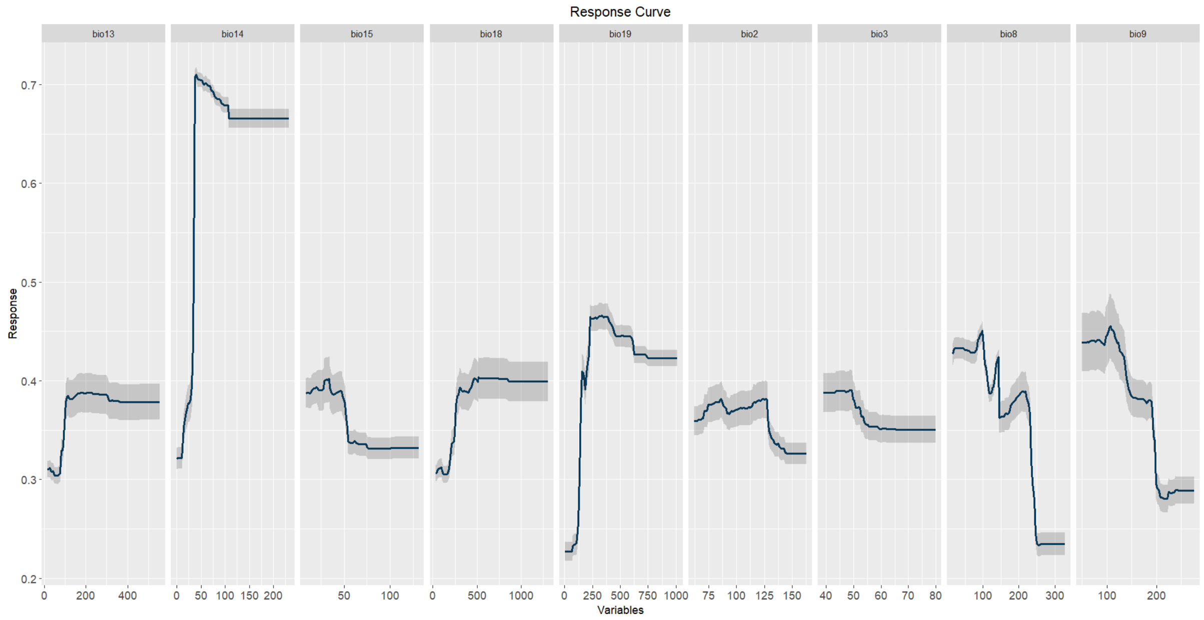
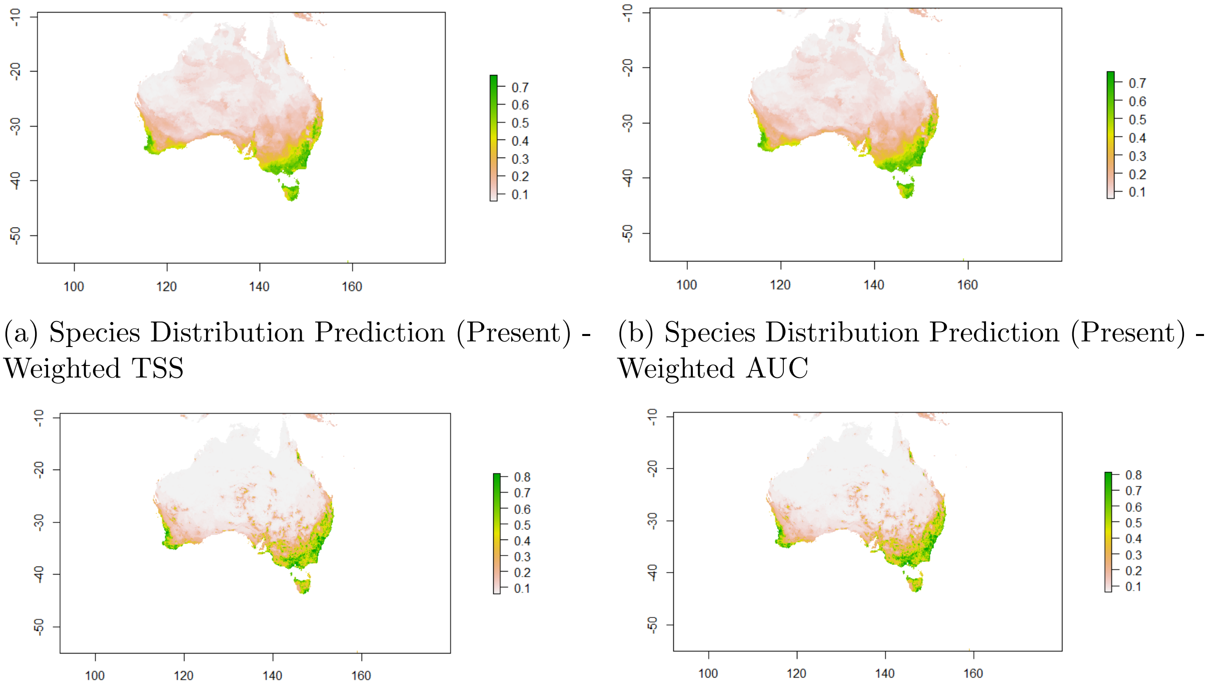


Figure 11: Response Curves - Ectotherms

5.4 Present and Future Species Distribution

Figure 12 and Figure 13 represent the distribution of endemic bird and frog species using historical and future climate data. The green parts indicate the highest probability of species occurrence and light shades of red indicate the lesser probability (ranging from 0.8 to 0.1). The prediction plots are generated using weighted ensemble techniques to choose the best prediction with TSS and AUC evaluation techniques.



(a) Species Distribution Prediction (Present) - Weighted TSS

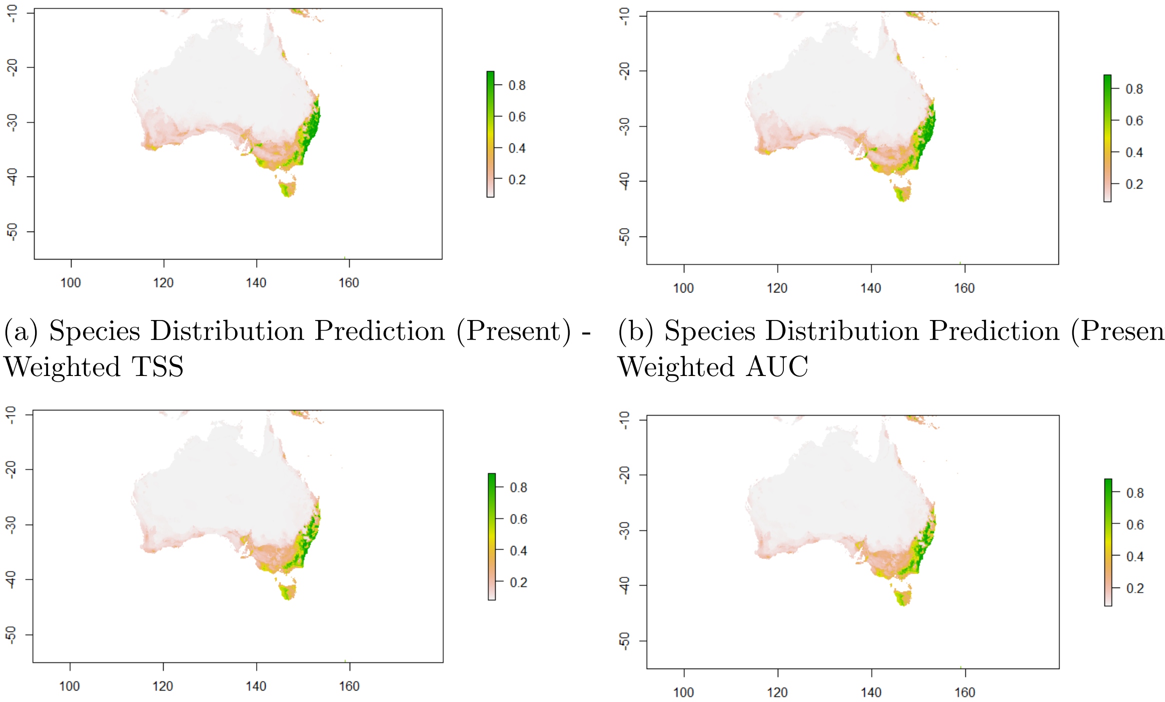
(b) Species Distribution Prediction (Present) - Weighted AUC

(c) Species Distribution Prediction (Future) - Weighted TSS

(d) Species Distribution Prediction (Future) - Weighted AUC

Figure 12: Distribution of Endothermic Birds Based on Present and Future Climate

It is clear from Figure 12a and Figure 12b that the species are abundantly in the Southern and Tasmanian Parts of Australia with 0.7 probability. Additionally, the endotherms are spread across the western and eastern regions having around 0.3 probability. Conversely, the decline of species can be clearly seen in the predictions generated using future climate variables in Figure 12c and Figure 12d. Nevertheless, the abundance areas remain the same as the present distributions with declining probabilities. The species distributed in the northern Australia are almost non-existent having only about 0.2 probability. The greener regions of Tasmania and southern Australia also seem to have receding occurrences.



(a) Species Distribution Prediction (Present) - Weighted TSS

(b) Species Distribution Prediction (Present) - Weighted AUC

(c) Species Distribution Prediction (Future) - Weighted TSS

(d) Species Distribution Prediction (Future) - Weighted AUC

Figure 13: Distribution of Ectothermic Frogs Based on Present and Future Climate

Figures 13a and 13b indicate the abundance of ectothermic frog species found to be more in the south-eastern regions of Australia having 0.6 to 0.8 probability. A small part of tasmanian regions also have sparse occurrences present. The future forecast indicated in Figure 13c and Figure 13d are not much different from the present predictions, then again, the western parts and some parts of south-eastern borders show declining habitats with occurrences reducing from 0.5 to 0.2 probability.

Overall, most occurrences of both endotherms and ectotherms are in the south-eastern and tasmanian regions. Comparitively, the receding occurrences of endotherms seem to be higher than the ectotherms in the next 70 years.

5.5 Discussion

Overall, the evaluation of the models under historic and current climate conditions indicates that the Random Forest with 5-Fold cross-validation performs better than the Boosted Regression Trees on the test data. This result holds true for both threshold-dependent and independent evaluation. However, the pseudo-absences created to compensate for the lack of absence sites in the data could have resulted in picking sites where the endemic species are actually present. Furthermore, the threshold-dependent and independent measures are seen to be dependent on the sample size while comparing the species group within the ectotherms and while comparing the endotherms and the ectotherms. The general trend that is followed in all the cases is that the models perform better on endotherms than on ectotherms due to the larger occurrence count of the former (depicted in Figure 5).

As a result of warming temperatures in the colder months, the distribution of endothermic species and ectothermic species in Australia is declining. This is indicated by the strong correlation between increasing precipitation, declining temperature during the colder months, and high evaluation metrics.

More importantly, present species distribution and future forecast (70 years ahead) with the carbon emission projected at RCP 8.5 ppm are compared for both endemics. It is found that the number of endotherms will plummet relative to the ectotherms. It can also be inferred that the endotherms are more affected by the carbon emission rates of the future than the ectotherms.

6 Conclusion and Future Work

Using Machine Learning Techniques and Statistical Tests, this research aimed to estimate the distribution of endemic species using empirical models, and to identify how climatic variables affect these models. In addition, this study also compares the distribution behaviors of ectothermic frogs and endothermic birds in relation to current and future climate variables.

The models are successfully generated using spatial data following a unique framework and using the newer *sdm* package in R.

This study can be further expanded by including the reproduction patterns, behaviours and other environmental variables such as soil type and vegetation types for each species group as explanatory variables. If more of such data is collected, reliable results can be derived from the robust machine learning techniques. Also, as Random Forest shows promising results it can be made more sophisticated by using Out-Of-Bag with bootstrapping techniques.

References

Bagariaa, P., Thapaa, A., Sharmaa, L. K., Joshia, B. D., Singha, H., Sharmaa, C. M., Sarmaa, J., Thakura, M. and Chandra, K. (2021). Distribution modelling and climate change risk assessment strategy for rare himalayan galliformes species using archetypal data abundant cohorts for adaptation planning, *Climate Risk Management, Science Direct* **31**: 100264.

Beery, S., Cole, E., Parker, J., Perona, P. and Winner, K. (2021). Species distribution modeling for machine learning practitioners: A review, *Proceedings of 2021 4th ACM SIGCAS Conference on Computing and Sustainable Societies, COMPASS 2021* pp. 329–348.

De Marco, P. and Nobrega, C. C. (2018). Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation., *PloS one* **13**(9): e0202403.

Dodangeh, E., Choubin, B., Eigdir, A. N., Nabipour, N., Panahi, M., Shamshirband, S. and Mosavi, A. (2020). Integrated machine learning methods with resampling algorithms for flood susceptibility prediction, *Science of the Total Environment* **705**.

Dyderski, M. K., Paz, S., Frelich, L. E. and Jagodzinski, A. M. (2018). How much does climate change threaten european forest tree species distributions?, *Global Change Biology, Wiley Online Library* **24**(3): 1150–1163.

Foden, W. B., Young, B. E., Akçakaya, H. R., Garcia, R. A., Hoffmann, A. A., Stein, B. A., Thomas, C. D., Wheatley, C. J., Bickford, D., Carr, J. A., Hole, D. G., Martin, T. G., Pacifici, M., Pearce-Higgins, J. W., Platts, P. J., Visconti, P., Watson, J. E. M. and Huntley, B. (2019). Climate change vulnerability assessment of species, *WIREs Climate Change* **10**(1): e551.
URL: *https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcc.551*

Fourcade, Y., Besnard, A. G. and Secondi, J. (2018). Paintings predict the distribution of species, or challenge of selecting environmental predictors and evaluation statistics, *Global Ecology and Biogeography* **27**(2): 245–256.

Gonzalez, S. (2018). Evaluating future impacts of climate change on traditional mexican maize suitablity and indigienous communities in mexico.

Hart, C. J., Kelly, R. P. and Pearson, S. F. (2018). Will the california current lose its nesting tufted puffins?, *PeerJ* **2018**(3): e4519.

Hoffmann, A. A., Rymer, P. D., Byrne, M., Ruthrof, K. X., Whinam, J., McGeoch, M., Bergstrom, D. M., Guerin, G. R., Sparrow, B., Joseph, L., Hill, S. J., Andrew, N. R., Camac, J., Bell, N., Riegler, M., Gardner, J. L. and Williams, S. E. (2019). Impacts of recent climate change on terrestrial flora and fauna: Some emerging australian examples, *A Journal of Ecology in Southern Hemisphere* **44**: 3–27.

Jiménez, L. and Soberón, J. (2020). Leaving the area under the receiving operating characteristic curve behind: An evaluation method for species distribution modelling applications based on presence-only data, *Methods in Ecology and Evolution* **11**(12): 1571–1586.
URL: *https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13479*

Kumar, L., Shabani, F. and Ahmadi, M. (2018). Assessing accuracy methods of species distribution models: Auc, specificity, sensitivity and the true skill statistic assessing accuracy methods of species distribution models: Auc, specificity, sensitivity and the true skill statistic assessing accuracy methods of species distribution models: Auc, specificity, sensitivity and the true skill statistic, *Global Journals Inc.* **18**(1).

Liu, C., Newell, G. and White, M. (2018). The effect of sample size on the accuracy of species distribution models: considering both presences and pseudo-absences or background sites, *Ecography, John Wiley Sons, Ltd* **42**(3): 535–548.

Liu, C., White, M. and Newell, G. (2009). Measuring the accuracy of species distribution models: a review, *18 th World IMACS / MODSIM Congress* pp. 13–17.

Naimi, B. and Araújo, M. B. (2016). sdm: a reproducible and extensible r platform for species distribution modelling., *Ecography, John Wiley Sons, Ltd* **39**: 368–375.

Shabani, F., Kumar, L. and Ahmadi, M. (2018). Assessing accuracy methods of species distribution models: Auc, specificity, sensitivity and the true skill statistic, *Acta Scientiarum Human and Social Sciences* **18**.

SUNG, S.-Y., LEE, D.-K., PARK, C., KIM, H.-G., KIL, S.-H., CHAE, H.-M., PARK, G.-S. and OHGA, S. (2018). Asessing effective sampling method and sample size for species distribution modeling of korean red pine (pinus densiflora), *Faculty of Agriculture, Kyushu University* **63**(2): 211–221.

Tiago, P., Pereira, H. M. and Capinha, C. (2017). Using citizen science data to estimate climatic niches and species distributions, *Basic and Applied Ecology* **20**: 75–85.

Valavi, R., Elith, J., Lahoz-Monfort, J. J. and Guillera-Arroita, G. (2019). blockcv: an r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models, *Methods in Ecology and Evolution* **10**(2): 225–232.