Implementing best practices and a workflow for modelling the geospatial distribution of migratory species

F. Santana [a], W. Hallgren [b], J. Rehn [a], L. Chiu [a], and H. Holewa [c]

[a] *Faculty of Education, Science, Technology & Maths, University of Canberra, Australian Capital Territory*
[b] *Griffith Climate Change Response Program, Griffith University, Gold Coast, Queensland*
[c] *Atlas of Living Australia, CSIRO, Australian Capital Territory*
Email: Fabiana.Santana@canberra.edu.au

Abstract: Species distributions are mainly determined by abiotic conditions, other species with which they interact, and the potential for dispersal and colonisation. Most of these factors are dynamic and change over time. Species' responses are also dynamic and may vary from evolved ecological niches to geographical isolation and speciation, to extinction.

In the past, summarising evolutionary processes into mathematical models would result in problems which were too complex and analytically intractable. A simplified modelling technique was developed to estimate geospatial species distributions with computationally viable solutions. It combines species presence/absence data with the relevant environmental layers for its survival (e.g., temperature and rainfall), calculating the species probabilistic distribution by applying machine learning and statistical algorithms. The resulting model is the species potential niche distribution, which can be applied to determine potential geographical areas for conservation and sustainable use of the environment, and to evaluate impacts of climate change, among other relevant applications.

Nowadays, supercomputers, big data, and cloud computing can be used to improve ecological modelling, but the research community is only gradually understanding the potential of these technologies. Researchers often use their own modelling environment (e.g., R/R Studio) and ignore the potential outcomes that can be brought about by large-scale computing and data resources. Modelling the geospatial distribution of migratory species, for example, requires running several models for different periods of time along a year, depending on the species migratory patterns, so it is a clear case for applying these new technologies.

This paper introduces a workflow to run cloud-based migratory species modelling. It determines the required steps to produce reliable models and best practices to properly design, understand, and evaluate such models. The workflow was implemented in the Biodiversity and Climate Change Virtual Laboratory (BCCVL), which is available for the research community. Experiments with the monthly and seasonal distribution of the migratory species *Danaus plexippus* (Monarch Butterfly) were conducted to assess the workflow and the BCCVL implementation. The generated models were compatible to results available in the literature, and they also matched the corresponding species data available in the most relevant species data portals worldwide, such as the Global Biodiversity Information Facility and the Atlas of Living Australia.

This workflow is the first step in a series of dynamic features that can be proposed to improve the current state-of-art in species distribution modelling with the help of new technologies such as cloud computing, HPC (High Performance Computing), and IoT (Internet of Things). Combined, they have the potential to take ecological niche modelling to the next level in terms of usability, availability, scalability, performance, and accuracy of the generated models.

Keywords: *Migratory species distribution, cloud computing, workflow, HPC, IoT, species distribution modelling, BCCVL (Biodiversity and Climate Change Virtual Laboratory)*

1. INTRODUCTION

Ecological modelling, including species distribution modelling, can be significantly improved by using supercomputers, big data, and cloud computing. These new technologies bring about the possibility of incorporating large-scale computing and data resources to the existing modelling techniques, allowing the development of faster models, reliable infrastructures with shared resources, and even more accurate solutions as the computational barriers to add more complexity to the current techniques disappear.

A widely applied technique to determine a species geospatial distribution is commonly known as ecological niche modelling, which combines species data (occurrence and absence, if available) with environmental data layers to determine a species probabilistic distribution. The technique applies machine learning and statistical algorithms, among others, to generate a model (Peterson, 2001; Santana et al., 2008). It is based on the hypothesis that, if a species is found under certain environmental conditions (*habitat*), then those conditions are the ones the species need to survive and reproduce.

Such models do not necessarily represent a species realised niche, as the latter may be determined by abiotic conditions, other species with which it interacts, the potential for dispersal and colonisation, and other dynamic factors that change over time (Barve et al, 2011). Still, as they show the species potential niche distribution projected onto the study area for a certain period of time, they have many applications. For example, ecological niche modelling has been applied to determine areas for conservation/sustainable use of the environment, and impacts of climate change on the biodiversity of a region (Santana et al., 2008).

This technique can be applied for modelling the geospatial distribution of migratory species, but in this case the data input to each model must correspond to a species migratory pattern. For example, an experiment with a species that has a seasonal migratory pattern will require a different snapshot for each season of the year. Running several species distribution model experiments will be required, which makes this a clear case for applying new technologies such as cloud computing and HPC (High Performance Computing).

This paper introduces a workflow to run cloud-based migratory species modelling, introducing the main steps to produce reliable models and best practices to properly design, understand, and evaluate such models. The proposed workflow was the basis for the implementation of the migratory species experiment in the Biodiversity and Climate Change Virtual Laboratory, BCCVL (Hallgren et al., 2016), where it is available for the research community. The BCCVL was supported by the National eResearch Tools and Resources Project (NeCTAR), an initiative of the Commonwealth being conducted as part of the Super Science Initiative and financed from the Education Investment Fund, Department of Industry, Innovation, Science, Research and Tertiary Education, Australia.

A case study was developed with the seasonal distribution of the migratory species *Danaus plexippus* (Monarch Butterfly) to assess the workflow and the accuracy of the BCCVL implementation. The results were compatible with the ones available in the literature.

This workflow was implemented as the first step in a series of dynamic features that can improve the current state-of-art in species distribution modelling by using cloud computing, HPC, and IoT (Internet of Things), among other new technologies.

2. WORKFLOW: MODELLING THE GEOSPATIAL DISTRIBUTION OF MIGRATORY SPECIES

This section introduces a workflow to describe the main steps required to obtain the geospatial distribution of a migratory species based on the ecological niche modelling technique. First, the technique itself is presented to illustrate how a single species distribution model can be obtained, for a specific point in time. Then, the workflow presents the specific requirements for the distribution of migratory species. The workflow was adapted from the reference process for ecological niche modelling presented in Santana et al. (2008), and it helps to understand the challenges associated with modelling migratory species distribution.

2.1. Obtaining a species distribution modelling based on the ecological niche modelling technique

Ecological niche modelling applies machine learning, statistical regression, and other analytics techniques to combine species presence and absence data (if available) with environmental variables such as temperature and rainfall, in order to calculate the species fundamental niche or habitat, which can be projected onto the study area to determine a potential species probabilistic distribution. Figure 1, adapted from Santana et al. (2008), details the main steps of the ecological niche modelling technique.

Initially (step 1), the researcher must define the experiment to be conducted, e.g., calculate the geospatial distribution of the Monarch Butterfly in Australia using seasonal environmental data.

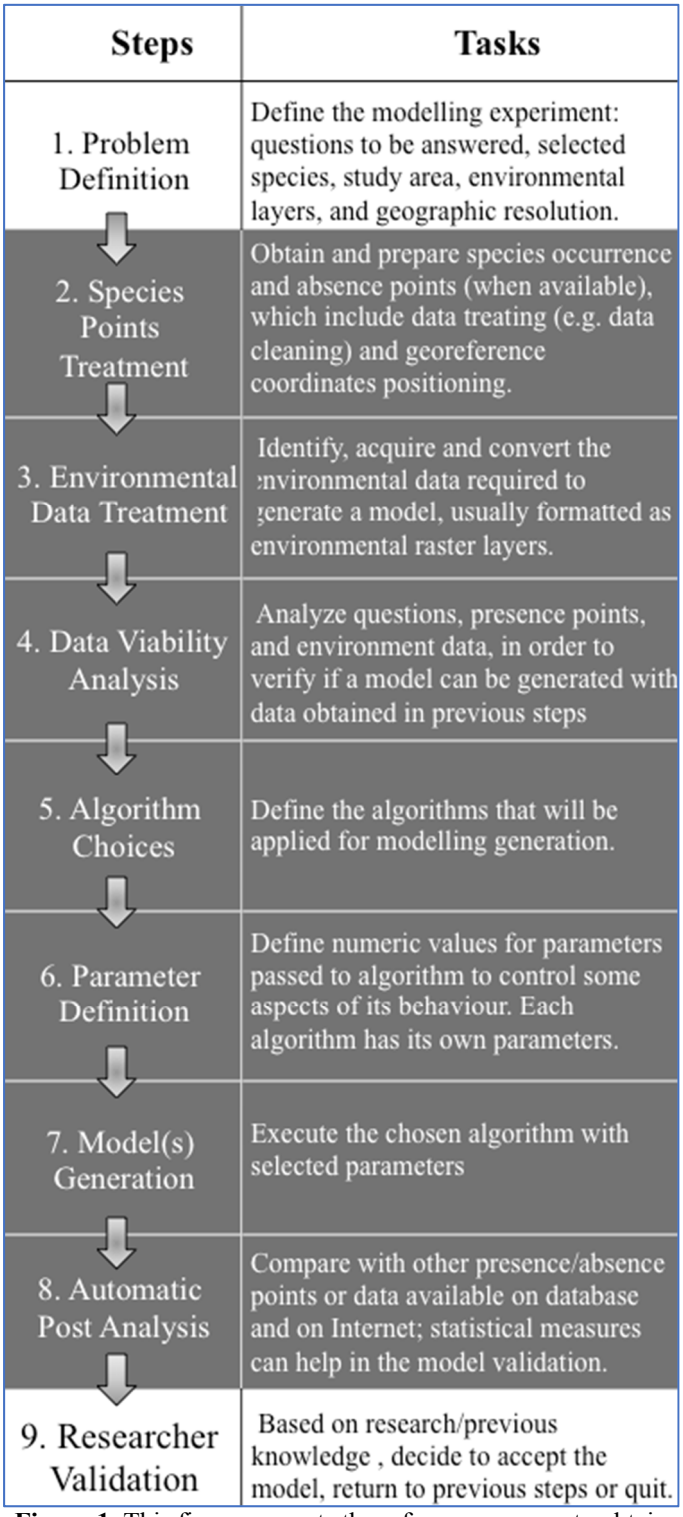

Figure 1. This figure presents the reference process to obtain the geospatial distribution of a species based on the ecological niche modelling technique –adapted from Santana et al. (2008), updated to reflect the current state of art of this technique.

In step 2, species occurrences must be obtained, cleaned, and organised in a proper format to allow the experiment to be conducted. This can be done manually or by using specific features available in the ecological niche modelling tools, such as R/RStudio [https://www.r-project.org/ and https://www.rstudio.com/], and openModeller [http://openmodeller.sourceforge.net/]. The researcher may use his or her own data, if available, or import datasets from species portals such as the Global Biodiversity Information Facility (GBIF) [http://www.gbif.org/] and Atlas of Living Australia (ALA) [http://www.ala.org.au/].

In step 3, the researcher must determine the environmental data layers that are relevant to that species survival. They should correspond to the layers that describe the known conditions of the species natural habitat, such as temperature, precipitation, and soil. E.g., if a species' reproduction cycle is affected by extreme temperatures, then layers representing the temperature of the coldest and hottest months should be added to the experiment. This data can be obtained from several sources, such as MODIS-FPAR[https://modis.gsfc.nasa.gov/data/dataprod/mod15.php] and Worldclim [http://www.worldclim.org/].

If the data obtained in steps 2 and 3 is adequate to generate a model, i.e., it is representative of a species and its habitat, then step 4 may progress. The viability analysis should assess if all requirements for generating a quality model have been addressed.

In step 5, one or more algorithms to generate a model should be chosen. Several types of algorithms are available, such as profile models, statistical regression models, machine learning models, and geographic models. It is not unusual to choose a number of different algorithms simultaneously (if the modelling tool allows this feature) and generate several models to compare the results. Depending on the software system architecture of the modelling solution, parallelisation may be applied. Most algorithms have parameters and they are also very relevant to the model generation, so those must be chosen and calibrated in step 6.

Once a model is generated in step 7, statistical analyses (e.g., AUC - area under curve), are calculated to help evaluate the model accuracy (step 8). Projection of the model onto the study area is part of this step. Finally (step 9), the researcher must analyse the final results and decide if a suitable model has been achieved or if further model generation/calibration is needed. In the latter, the researcher will return to a previous step of the process (e.g., step 6), make the required adjustments (e.g., redefine the algorithm's parameters), and rerun the model.

2.2. Obtaining the distribution of migratory species based on the ecological niche modelling technique

An ecological niche model presents a species potential distribution projected onto the area of interest at a specific point in time. The distribution of a migratory species varies according to its migratory patterns, so the generation and combination of multiple models, obtained at different periods of time and potentially projected onto different distributional areas (or a larger area incorporating the whole spectrum of the species potential movements), is required to correctly describe and understand the species migratory movements. For example, if the intention is to analyse the behaviour of a migratory species along the four seasons of the year (Summer, Autumn, Winter and Spring), then four different ecological niche models, one per season, are required. If the intention is to analyse the behaviour of such a species along the twelve months of the year, from January to December, then twelve experiments are necessary, one per month.

The process is similar to the one presented in Figure 1, however:

- The species data obtained in step 2 must be organised in accordance with the migratory pattern to be studied, i.e., if the researcher is modelling a species accordingly to a seasonal migratory pattern, different datasets must be organised for the four seasons of a year;
- The environmental data (step 3) must be carefully analysed to understand the type of data required (e.g., monthly, seasonal datasets) and, if it varies according to a species migratory pattern, then different datasets must be organised for each experiment; again, if a species is being modelled according to a seasonal migratory pattern and temperature is a relevant environmental layer, then a temperature layer for each season will be necessary –this is easily understandable, as temperatures in Summer and Winter may vary considerably in certain regions of the world;
- The model generation and automatic post analysis (step 7 and 8) will require the generation of not one, but several different models.

An important issue mentioned in section 2.1 is the projection area (step 7). As the model is originally generated in the subspace of ecological niche conditions, the projection area is usually defined as part of the experiment by the researcher (e.g., modelling a species distribution for Australia, or for the Australian Capital Territory, or for a specific biogeographic region). For migratory species, special consideration to this aspect must be given based on the potential areas where the species have been observed, as restricting the projection area may prevent a proper study of the species distribution.

3. A CASE STUDY: MODELLING THE MONARCH BUTTERFLY WITH THE BIODIVERSITY AND CLIMATE CHANGE VIRTUAL LABORATORY (BCCVL)

In order to assess the proposed workflow for obtaining the distribution of migratory species based on the ecological niche modelling technique, a case study was developed with the *Danaus plexippus* (Monarch Butterfly) migratory species. The BCCVL implements different experiments related to species distribution modelling as workflows, including a solution to support migratory species distribution based on the process described in section 2. The BCCVL is cloud-based, so the researcher can run several experiments simultaneously and compare the results. The BCCVL thus was the chosen solution to run the case study.

Step 1 – Problem definition

The workflow starts by defining the experiment. A title must be chosen and a relevant description of the model to be generated must be provided. This corresponds to the step 1 in the process presented in Figure 1, so the researcher should take this opportunity to properly define the modelling experiment by identifying the questions to be answered, species to be studied, projection area, and environmental layers. In this case, the purpose of the experiment is to calculate the seasonal geospatial distribution of the Monarch Butterfly in Australia based on seasonal climate data. The resolution will be automatically adjusted by the BCCVL.

Step 2 – Species data treatment

In this step of the process, the researcher should prepare the species occurrence and absence points for the experiment. Providing occurrence data is mandatory for model generation, but the data may require specific treatment, such as data cleaning and georeferencing adjustments before being used. Many ecological niche modelling algorithms also require absence data, or at least pseudo-absence data. The researcher may provide his or her own absence datasets if they are available, or define a strategy for generating pseudo-absence datasets if one is not automatically provided by the algorithm or modelling tool.

The BCCVL has a feature that allow users to select the species occurrence datasets, which may be uploaded by the user, directly imported from the ALA, or reused from a list of previously imported datasets. Once the dataset is available in the BCCVL, it can be used in several different experiments. A species dataset can also be shared between researchers so as to allow collaboration, and it will also appear in the list of available datasets. In this case study, the *Danaus plexippus* data was manually obtained from the ALA and manually cleaned to remove spurious occurrences, then uploaded to the BCCVL. As the *Danaus plexippus* is a migratory species, the occurrences should also be organised according to the experiments that will be conducted, which in this case means the dataset should be split into four seasonal datasets corresponding to the four seasons of the year. This issue is already resolved by the BCCVL migratory species experiment, so manually organising the datasets was not necessary.

As absence datasets were not available, pseudo-absence data generation was required. The BCCVL implements a number of strategies generate pseudo-absences for modelling purposes, such as random, contrasting environment and min-max radius. For modelling the distribution of the *Danaus plexippus*, the random strategy was chosen.

Step 3 – Environmental data treatment

In this step, the researcher should identify, acquire and convert the environmental data required to generate a model, usually formatted as environmental raster layers.

The BCCVL has a number of relevant environmental layers available for modelling generation, which have already been prepared for modelling purposes. For the *Danaus plexippus* distribution, the ANUClim (Australia), Current Climate, (1976-2005), 30 arcsec (~1 km) (Hutchinson et al., 2014), monthly datasets were chosen. This data is organised in twelve datasets, each one containing the climate data for a specific month. As the case study was focused on seasonal distribution, four groups with three datasets were organised for the migratory species experiment, each group containing the datasets corresponding to a particular season (e.g., Summer in Australia corresponds to the months of December, January, and February).

The BCCVL also enables, at this point, the definition of geographic constraints for training the model, a requirement for many modelling algorithms. The convex hull option was selected for the case study.

Step 4 – Data viability analysis

The main purpose of this step is to decide if generating a model with the available data is a viable choice or not. The BCCVL migratory experiment will fail if incompatible data or parameters are provided, but the most important assessment to be made is unrelated to any modelling tools. For example, if the input species data and environmental data correspond to different timeframes, the results may not be accurate, and if a researcher does not really understand a species habitat and incorrectly selects the environmental layers for a species survival, the resulting model may be incorrect from the ecological/environmental point of view.

Steps 5 and 6 – Choosing algorithms and corresponding parameters to run the experiment

In this step, the researcher should select the algorithm (or set of algorithms) that will be applied for model generation. For each algorithm, a number of parameters will be available for tuning the model. Those should also be chosen in this step.

The modelling of the *Danaus plexippus* was generated with MaxEnt with the default parameters. MaxEnt is an algorithm that predicts species occurrences by finding the distribution that is most spread out, or closest to uniform, while taking into account the limits of the environmental variables of known locations (Phillips et al, 2006). It is also one of the most applied algorithms for species distribution modelling worldwide.

Steps 7, 8 and 9 – Model generation, automatic analysis and researcher validation

After running the experiment, the seasonal distribution of the migratory species *Danaus plexippus* was generated by the BCCVL, for each season (step 7). The results were combined using QGIS [http://www.qgis.org/] to easily show the species migratory movements, which are presented in Figure 2. Each map provides the original distribution of a species for a given season, the areas where no changes were observed, and the species migratory movements from one season to the next.

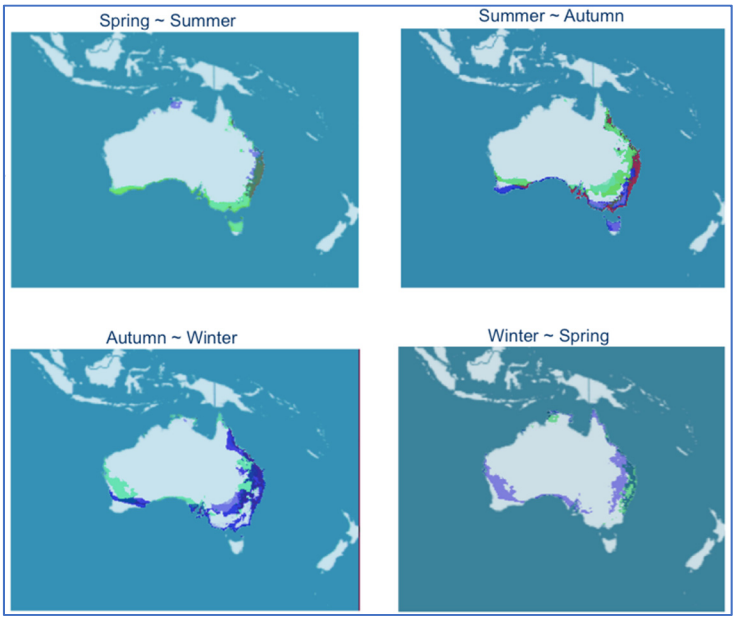

Figure 2. Seasonal distribution of the *Danaus plexippus* migratory species. The original distribution for each season is shown in green, red are the areas where no changes were observed, and blue represents the species migratory movements from one season to the next.

The BCCVL calculates several statistical analyses per model, to help the researcher to decide if the model is acceptable or if adjustments are required (step 8). For this case study, the average AUC = 0.96, which was a good indicator of model accuracy. As the resulting models of this case study are also compatible with others found in the literature for the *Danaus plexippus* distribution, such as those presented in Hoth et al. (1997), no further refinements were required and the model was deemed acceptable (step 9), concluding the case study.

4. DISCUSSION

Researchers have been successfully calculating species distribution models for years in their local computing environments (e.g., laptops and desktops). However, the application of cloud computing and HPC technology may represent a significant step forward in improving the current techniques. In this environment, quality attributes such as usability, availability, scalability, and performance can be easily improved, as well as easy access to previous developments such as algorithms, statistical evaluations and projection strategies. For example, in the migratory species experiment, the BCCVL has resolved many issues that otherwise would have to be manually resolved by the researcher (e.g. organising datasets), improving usability, and the distributed nature of its cloud-based implementation allows the generation of several experiments in parallel, improving scalability and performance, when compared to desktop solutions.

Another important aspect of using cloud-based services is the possibility of improving reproducible research and provenance. A cloud-based solution such as the BCCVL can easily be extended or integrated to another solution to allow model publication with the associate provenance elements (datasets, algorithms, parameters and any other configuration required to generate a model). The BCCVL already provides the provenance information along with a model, so it is just a matter of organising its publication in an appropriate manner.

A number of considerations must be made related to the input data for model generation. First, it is worth noting that ALA records, as well as the records provided by any citizen science-based species portal, mainly rely on the information provided by its contributors. This is the reason why we first cleaned the ALA datasets to remove spurious/impossible values in our case study (e.g., occurrences that do not correspond to the period of the experiment or impossible values that probably correspond to observations in zoos or artificial habitats). The environmental data, on the other hand, may be an important limiting factor for specific experiments, such as modelling migratory species. Most data provided by WorldClim (Hijmans et al., 2005) and others is organised in layers that present averages for long periods of time (e.g., aggregation for 5, 10 or even 30 years.) This was partially caused by the computational constraints of desktop solutions, where the researcher actually had to download the relevant datasets. With the usage of cloud-based solutions the computational restriction disappears, however most of the environmental data portals are still providing averages instead of non-aggregated data.

Further automation of the modelling process is possible and desirable, and there are many opportunities for doing so. For example, automatic data cleaning tools may be incorporated to either the species provider portal or to the species modelling workflow, allowing direct import of the datasets in all cases (step 2). Also, if a species habitat is well known and available along with the datasets, the selection of environmental layers (step 3) could be automated. Parameter selection is another good candidate for automation, however this may be quite complex. Depending on the problem, an AI (Artificial Intelligence)-based approach may be required.

Significant improvements can be made to the accuracy of migratory species modelling by associating ecological niche modelling techniques with remote sensing data (Fern et al., 2017), GPS-tracking devices, drones, etc. Technology for building such combined models already exists, however costs, integration and interoperability issues still represent significant challenges for the research community.

Finally, it is worth noting the workflow presented here is only the first step in a series of dynamic features that can be proposed to improve the current state-of-art in species distribution modelling with the help of new technologies such as cloud computing, HPC and IoT. Combined, they have the potential to take this technology to the next level.

5. CONCLUSION

This paper presented a workflow for migratory species distribution based on the ecological niche modelling technique, as well as a case study to help assess its viability and correctness, and a brief discussion on potential issues and future improvements. Further experiments are required to fully validate the workflow, but so far the results obtained are promising.

The workflow also shows clear advantages of using cloud-based tools such as the BCCVL for migratory species distribution, as they potentially facilitate modelling reuse, provenance, reproducible research and model publication.

ACKNOWLEDGMENTS

The authors thank NeCTAR for the support to the BCCVL project. The authors also thank the University of Canberra for supporting their participation in MODSIM 2017.

REFERENCES

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S.P., Peterson, A.T., Soberón, J. and Villalobos, F. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, 222(11), pp.1810-1819.

Fern, R.R. and Morrison, M.L. (2017). Mapping critical areas for migratory songbirds using a fusion of remote sensing and distributional modeling techniques. *Ecological Informatics*, 42, pp.55-60.

Hallgren W, Beaumont L, Bowness A, Chambers L, Graham E, Holewa H, Laffan S, Mackey B, Nix H, Price J and Vanderwal J. (2016). The Biodiversity and Climate Change Virtual Laboratory: Where ecology meets big data. *Environmental Modelling & Software*, 76, pp.182-186.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.,* 25: 1965–1978. doi:10.1002/joc.1276

Hoth, J., Merino, L., Oberhauser, K., Pisanty, I., Price, S. and Wilkinson, T. (1999). 1997 North American Conference on the Monarch Butterfly. *Commission for Environmental Co-operation*.

Hutchinson M, Kesteven J, Xu T (2014). Monthly climate data: ANUClimate 1.0, 0.01 degree, Australian Coverage, 1976-2005. Australian National University, Canberra, Australia. Made available by the Ecosystem Modelling and Scaling Infrastructure (eMAST, http://www.emast.org.au) of the Terrestrial Ecosystem Research Network (TERN, http://www.tern.org.au).

Peterson, A.T. (2001). Predicting Species' Geographic Distributions Based on Ecological Niche Modeling. *The Condor*, 103(3), pp.599-605.

Phillips, S.J., Anderson, R.P. and Schapire, R.E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3), pp.231-259.

Santana, F.S., De Siqueira, M.F., Saraiva, A.M. and Correa, P.L.P. (2008). A reference business process for ecological niche modelling. *Ecological informatics*, 3(1), pp.75-86.