The Biodiversity and Climate Change Virtual Laboratory: How Ecology and Big Data can be utilised in the fight against vector-borne diseases

Willow Hallgren [a], **Linda Beaumont** [b], **Andrew Bowness** [a], **Lynda Chambers** [c], **Erin Graham** [d], **Hamish Holewa** [a], **Shawn Laffan** [e], **Brendan Mackey** [a], **Henry Nix** [f], **Jeff Price** [g], **Jeremy Vanderwal** [d], **Rachel Warren** [g] **and Gerhard Weis** [a]

[a] *Griffith University, Gold Coast campus, Parklands Drive, Southport, QLD, Australia, 4215;* [b] *Department of Biological Sciences, Faculty of Science and Engineering, Macquarie University, NSW, Australia, 2109;* [c] *Bureau of Meteorology, Melbourne, Australia, 3001;* [d] *eResearch and the Centre for Tropical Biodiversity and Climate Change, James Cook University, Townsville, QLD, Australia, 4810;* [e] *Centre for Ecosystem Science, School of Biological, Earth and Environmental Science, University of New South Wales, Australia, 2052;* [f] *The Fenner School of Environment and Society, The Australian National University, Building 141 Linnaeus way Canberra, ACT, Australia, 2601;* [g] *Tyndall Centre for Climate Change Research, School of Environmental Sciences, University of East Anglia, Norwich, NR4 7TJ, United Kingdom.*
Email: w.hallgren@griffith.edu.au

Abstract: Advances in computing power and infrastructure, increases in the number and size of ecological and environmental datasets, and the number and type of data collection methods, are revolutionizing the field of Ecology. To integrate these advances, virtual laboratories offer a unique tool to facilitate, expedite, and accelerate research into the impacts of climate change on biodiversity. We introduce the uniquely cloud-based Biodiversity and Climate Change Virtual Laboratory (BCCVL), which provides access to numerous species distribution modelling tools; a large and growing collection of biological, climate, and other environmental datasets, as well as a variety of experiment types to conduct research into the impact of climate change on biodiversity. Users can upload and share datasets, potentially increasing collaboration and cross-fertilisation of ideas and innovation among the user community. Feedback confirms that the BCCVL's goals of lowering the technical requirements for species distribution modelling, and reducing time spent on such research, are being met. We present a case study that illustrates the utility of the BCCVL as a research tool that can be applied to the problem of vector borne diseases and the likelihood of climate change altering their future distribution across Australia. This case study presents the preliminary results of an ensemble modelling experiment which employs multiple (15) different species distribution modelling algorithms to model the distribution of one of the main mosquito vectors of the most common vector borne disease in Australia: Ross River Virus (RRV). We use the BCCVL to do future projection of these models with future climates based on two extreme emissions scenarios, for multiple years. Our results show a large range in both the modelled current distribution, and projected future distribution, of the mosquito species studied. Most models (that were built using different algorithms) show somewhat similar current distributions of the species however there are three models that are obvious outliers. The projected models show a similar range in the distribution of the species, with some models indicating a fewer areas (and also areas with a lower probability of occurrence in specific areas) where the species is likely to be found under a climate change scenario. However, a majority of models show an expanded distribution, with some areas that have a greater probability of the occurrence of this species; this will provide a more robust indication of future distribution for policy makers and planners, than if just one or a few models had been employed. The economic and human health impact of vector borne diseases underline the importance of scientifically sound projections of the future spread of common disease vectors such as mosquitos under various climate change scenarios. This is because such information is essential for policy–makers to identify vulnerable communities and to better manage outbreaks and potential epidemics of such diseases. The BCCVL has provided the means to effectively and robustly bracket multiple sources of uncertainty in the future spread of RRV: this study focuses on two of these - the future distribution of a primary mosquito vector of the disease under two extreme scenarios of climate change. Research is underway to expand our analysis to take into account more sources of uncertainty: more vector and amplifying host species, emissions scenarios, and future climate projections from a range of different global climate models.

Keywords: *Biodiversity, climate change, virtual laboratory, species distribution modelling*

1. INTRODUCTION AND PURPOSE

Many fields of research are undergoing a methodological revolution, and in recent years this has particularly applied to the fields of both ecology and "e-research", which is the application of Information and Communication Technologies (ICT), tools and infrastructure to scientific investigations. This is in part due to the emergence of 'Big Data', which we define as "a combination of an approach to informing decision making with analytical insight derived from data, and a set of enabling technologies that enable that insight to be economically derived from at times very large, diverse sources of data. Advances in Earth system and biological sensing technologies have created the opportunity to use large scale, fine grained data to understand natural systems and behaviours; while innovation in technology makes it viable economically to use that information to inform decisions and improve outcomes." (adapted from J. Akred at https://datascience.berkeley.edu/what-is-big-data).

There has been a huge growth in the size, complexity and diversity of datasets, as well as computational and analytical improvements in statistical and simulation models (such as machine learning: Peters et al*.,* 2014) in recent years. For example, data repositories such as the Atlas of Living Australia (ALA, http://www.ala.org.au) and the Global Biodiversity Information Facility (GBIF, http://www.gbif.org/) have 50 and 530 million specimen records respectively, and are increasing the rate at which they amass information. Satellite and airborne sensors are also generating petabytes of spatially explicit environmental data and increasing the diversity of available data types. These developments have enabled more complex SDM experiments; we argue that while many individual SDMs may feasibly be run using traditional processor capability, when multiple SDMs are run, with large numbers of often complex data inputs, under multiple climate change scenarios generated with multiple climate models, and also under different assumptions (and settings), this ensembling approach has the potential to propel SDM into the realms of "Big Data".

e-Research has the potential to facilitate research and the creation of new scientific insights in the field of Ecology, due to technological advances which have underpinned the unprecedented and growing ability to securely store, manage, share, analyse and synthesise research data within and across disciplines, This combination of e-Research and Big Data has made possible the development of the Biodiversity and Climate Change Virtual Laboratory (BCCVL). The BCCVL is a comprehensive platform for species distribution and trait modelling, and is designed to assist the ecological research community by connecting researchers to existing and new research facilities; datasets, data repositories, and major data storage and management facilities; and the high-performance computational, analytical, work-flow and visualisation tools enabled by e-Research. Through a cloud-based e-Research facility, the BCCVL provides researchers, environmental managers, policy analysts and other interested communities with access to:

- a suite of the most commonly used and robust modelling tools and functions to spatially analyse biological data;
- a comprehensive set of climate change data comprising monthly estimates of downscaled climate change projections;
- national extent ancillary physical, environmental, vegetation and land cover data;
- important post-modelling diagnostic, mapping and other visualization capacities;
- a facility to upload and share data and workflows; and
- the means to undertake spatial modelling at multiple spatial scales down to 250 m resolution for the entire country.

The BCCVL is currently populated with predominantly Australian datasets, but it has not been specifically customised for Australia. The BCCVL can assimilate user data and therefore be used to do species distribution modelling and other biodiversity analyses anywhere in the world: current research using the BCCVL has expanded to the analysis of species in other countries, confirming the BCCVL's usefulness in conducting biodiversity research internationally. Moreover, the BCCVL can also serve as a template for other countries wanting to set up their own Virtual Laboratory.

There are many advantages to using the BCCVL: it enables researchers to conduct modelling experiments and related analyses far more efficiently and effectively. It decreases the preparation time associated with species distribution modelling. This includes data reparation which can involve identifying applicable data (searching, defining domain, defining dataset assumptions and limitations), acquiring that data (involving licencing issues, acquiring rights statements, downloading/buying/retrieving the data), scaling/standarising the data (to ensure consistency with other datasets in the model, formatting data for model ingestion and visualisation), and then validating the data for suitability for each algorithm, and visualisation). Without the BCCVL you would then need to set up the modelling environment (which may involve learning the R

language (which can take from hours to many weeks), importing data into R, identifying the algorithm/s to use and the R package required for each algorithm, importing R package/s into the R environment and then running the algorithm/s (all individually), and then learn how to visualise and manipulate outputs (to create maps/graphics). Depending on the individual user, this process could take from weeks to months. The BCCVL negates the need for any of this preparatory work - it also negates the need for script writing, advanced programming or modelling expertise.

The resulting increase in research capability and efficiency will facilitate the development of additional research trajectories currently not possible due to logistical and the computational limitations of individuals and many research groups. For example, ensemble modelling experiments involving large numbers of SDMs, climate model projections, and emissions scenarios, which were once logistically quite challenging, are now accessible via a web browser.

The BCCVL is likely to set a new standard in the complexity and comprehensiveness of the experimental design expected of species distribution modellers. For example, the ability to perform large numbers of SDM experiments easily will enable more comprehensive comparisons of different SDMs, competing sets of potential explanatory variables, the available climate change models, emissions scenarios and projection periods. It will also greatly facilitate sensitivity analyses on parameter values.

Using the BCCVL will increase research productivity (in terms of time saved and scientific output), which will confer advantages both at the level of the individual scientist and research communities. Moreover, the BCCVL enables researchers to share data and modelling frameworks, promoting the use and reuse of data, which is currently underexploited (Peters et al. 2014) and enabling greater transparency in the research process.

2. DESCRIPTION OF THE BIODIVERSITY AND CLIMATE CHANGE VIRTUAL LABORATORY

A link from the BCCVL homepage at http://www.bccvl.org.au allows anyone with an Australian Access Federation (AAF) password to log into the BCCVL. Other domestic and international users can request a login account with the BCCVL, or log in either by acquiring a guest AAF account, or via the AAF Virtual Home.

2.1. Structure and functionality

The BCCVL comprises three components:

Datasets

The *Datasets* section of the BCCVL houses species location and trait data, current and future climate data, and other environmental data (e.g., soil, geology and vegetation type). Brief dataset summaries are listed on the front page of the dataset section, which provides the choice of viewing a map of the dataset (overlain on a national map), downloading the dataset, or accessing metadata in a pop-up box. This page allows users to search among the datasets provided by the BCCVL, shared datasets, and self-uploaded datasets.

Searches can be filtered by dataset type (species absence, abundance, occurrence, and traits; current and future climate, and other environmental datasets), as well as resolution (90 m, 250 m, 1 km, 5 km, 10 km, 20 km, and 50 km). The datasets page also provides a facility to search for, view, import and share a species dataset from the online repositories such as the ALA. Users can upload their own species occurrence, abundance or trait datasets, and other environmental datasets, and share them with fellow BCCVL users if they choose to. Recent changes to open (free) data policies, such as the Landsat program, and the requirements of some government funding bodies, have greatly expanded the range of data accessible to researchers. A list of datasets currently accessible within the BCCVL are listed in Hallgren et al. (2015), and the BCCVL has the capacity to add additional datasets as needed.

Climate Change Projection Data: The future climate data that the BCCVL provides for projecting species distribution models with climate change scenarios, is derived from global climate change modelling which was carried out by using the MAGICC4.1 climate model (Wigley and Raper, 2001; Lowe et al., 2009) and the pattern-scaling module ClimGEN (developed from Mitchell, 2003; see also Warren et al., 2008; Osborn, 2009), both of which are designed explicitly to emulate the behaviour of the complex models so that impact modelers can study emission scenarios that have not been simulated by the GCM modelers. Hallgren et al. (2015) describes in detail the methodology used to produce the climate change projection data utilised in the BCCVL.

Experiments

The Experiments section of the BCCVL allows users to access a suite of statistical modelling and analytical tools. There are currently five different types of experiments users can undertake:

a) *Species Distribution Modelling Experiments* identify the potential distribution of a species given current climate conditions;

b) *Climate Change Experiments* project a current species distribution into the future based on a climate projection, for one or more emission scenarios;

c) *Biodiverse Experiments* calculate biodiversity statistics (species richness, rarity and endemism) based on species distribution modelling results;

d) *Species Trait Modelling Experiments* (STMs) identify future distributions of a particular species trait (e.g. Leaf Area Index);

e) *Ensemble Modelling Experiments* enable the utilization of multiple models (SDMs, STMs or climate models) or scenarios to reduce some of the uncertainty inherent in the single-model/scenario approach.

The SDMs are the core functionality of the BCCVL as their results are used in most of the other components. There are currently 19 SDM algorithms available, as well as five algorithms employed in species trait modelling. These include the popular and well-known MaxEnt (Phillips et al. 2006) and Artificial Neural Networks (Hilbert,& Van Den Muyzenberg,1999), as well as simpler and more easily comprehensible algorithms, such as Bioclim (Nix, 1986) and Mahalanobis Distance (Mahalanobis, 1936). A demonstration of how to use the BCCVL to implement a species distribution modelling experiment, as well as a list of all algorithms currently employed in the BCCVL is given on homepage of the BCCVL website (http://www.bccvl.org.au).

The algorithms that are currently implemented in species trait modelling within the BCCVL include not only algorithms specifically designed to model species distributions (e.g. MaxEnt, Bioclim), but also widely used statistical methodologies such as Generalized Additive Models, Generalized Linear Models, and also other common statistical methodologies such as linear models, analysis of variance, and multivariate analysis of variance. The BCCVL automatically facilitates modelling experiments at multiple scales: currently the range of resolutions available for modelling experiments is 90 m to 50 km.

Knowledge Base and Decision Support Tool
The Knowledge Base is designed as a repository of information about many facets of the BCCVL. It includes a glossary, background information on all modelling algorithms, links to key references and papers, as well a Decision Support Tool to guide the user, and provide as much information as they need to conduct the experiments offered by the BCCVL in an intelligent and sensible manner. This feature of the BCCVL will also be improved by continuing user input and feedback.

2.2. Technical details
The BCCVL utilises a variety of open source software packages. These are operated on the Australian National eResearch Collaboration Tools and Resources (NeCTAR) Research Cloud, which will provide 35,000 cores of processing capacity hosted at eight nodes (data centres) distributed across Australia. The BCCVL's architecture is designed to handle large datasets, process data through experiments, display experiment outputs and securely share data within a cloud-based setting. The BCCVL is novel in its utilisation of cloud-based technologies to perform modelling functions traditionally reserved for cluster services or purpose built High Performance Computing. Cloud based technologies are built on enterprise level data centre infrastructure and achieve scale and resilience through the ability to easily add and replace individual components; they will enable the BCCVL to easily scale to meet new demands for processing or storage capacity.

Within the application, the BCCVL is composed of six discrete components, comprising: (i) visualizer, (ii) front-end user interface, (iii) back-end manager, and (iv) data mover components, as well as (v) job execution and worker node, and (vi) swift object storage components. These components communicate through common Application Programming Interfaces (APIs) such as SOAP, JSON and XL-RPC to enable modularity and the ability to add additional resources or features to the BCCVL whilst in operation. All code is open source and available on GitHub at https://github.com/BCCVL. The biodiversity experiments are implemented using the Biodiverse platform (Laffan et al. 2010; http://purl.org/biodiverse). Hallgren et al. (2015) provides a schematic of the major components and information architecture of the BCCVL, as well as further technical details on the six components that constitute the BCCVL.

3.   CASE STUDY: USING THE BCCVL TO MODEL CURRENT AND FUTURE DISTRIBUTION OF A MAJOR DISEASE VECTOR SPECIES
We present here a case study that illustrates the utility of the BCCVL as a research tool that can be applied to the problem of vector borne diseases and the likelihood of climate change altering their future distribution across Australia. This case study presents the preliminary results (Fig. 1) of an ensemble modelling

experiment which employs multiple (15) different species distribution modelling algorithms to model the distribution across Australia of one of the main mosquito vectors of the most common vector borne disease in Australia: Ross River Virus (RRV).
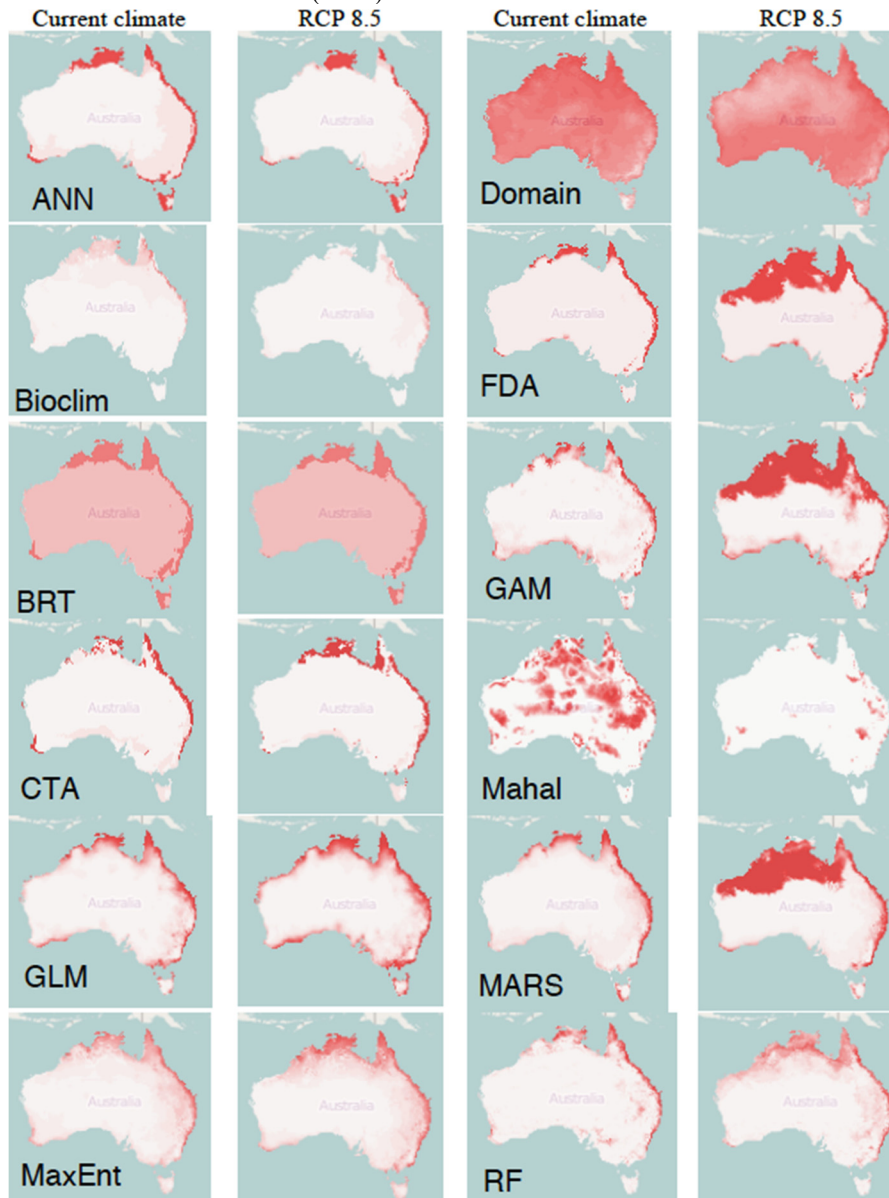

Figure 1. Results of 12 Species Distribution Models run with current climate and projected climate for 2085, for the Representative Concentration Pathway 8.5 scenario. The 12 models are Artificial Neural Network (ANN), Bioclim, Boosted Regression Tree (BRT) model, Classification Tree Analysis (CTA), Generalized Linear Model (GLM), Maximum Entropy Model (MaxEnt), Domain, Flexible Discriminant Analysis (FDA), Generalized Additive Model (GAM), Mahalanobis Model (Mahal), Multiple Adaptive Regression Splines (MARS), and Random Forest (RF).

A dataset of 932 occurrences (comprising latitude/longitude coordinates) of the mosquito species *Aedes vigilax* was downloaded from the GBIF website and uploaded to the BCCVL. We then ran a 'Species Distribution Modelling Experiment' in the BCCVL with this data, which comprised six predictor variables for current climatic conditions which are known to influence the distribution of this species; annual mean temperature, maximum temperature of the warmest month, minimum temperature of the coldest month, annual precipitation, precipitation of the wettest month and precipitation of the driest month, all having a 5 km resolution) (Yu et al., 2014).

These variables were input into 15 species distribution modelling algorithms that have been incorporated into the BCCVL; Artificial Neural Network (ANN), Bioclim, Boosted Regression Tree (BRT), Circles,

Classification Tree (CT), Convex Hull (CH), Domain, Flexible Discriminant Analysis (FDA), Generalized Additive Model (GAM), Generalized Linear Model (GLM), Mahalanobis Model (Mahal), Maximum Entropy Modeling (Maxent), Multiple Adaptive Regression Splines (MARS), Random Forest, (RF) Voronoi Hull Model (VH)).

We then ran a 'Climate Change Experiment' in the BCCVL to model the future projection of *A. vigilax* (using the same predictor variables) with multiple future climates based on two extreme emissions scenarios (RCP2.6, RCP8.5, as simulated by the CSIRO Mark3 global climate model). We ran this experiment for two points in time representing the medium (2045) and long term (2085) future.

We present here the results for only 12 out of the 15 SDMs, due to the divergence of the projected distributions of the Circles, VH and CH models from those of the rest of the 15. Due to these three models being mainly used for pedagogic purposes in the current SDM literature, we chose to exclude these from this analysis and focus on the models with known better predictive skill.

We also only present here the Climate Change Experiment results for 2085, due to the much smaller differences seen between the modelled distributions based on the current climate and the year 2045 climate projection. Our results show a large modelled range both within and between the current and projected future distribution of *A. vigilax*. Most models show somewhat similar distributions of the species however there are three models that are obvious outliers (Domain, Mahal and BRT) (Fig.1).

The projected models show a similar range in the distribution of the species, with four models indicating fewer areas (and often a lower probability of occurrence in specific areas) where *A. vigilax* is likely to be found under the RCP8.5 climate change scenario in 2085 (ANN, Bioclim, BRT, and Mahal). Domain shows a slight decrease in the probability of occurrence over large areas of Australia. However, seven models CTA, FDA, GLM, GAM, Maxent, MARS, and RF) show an expanded distribution of *A. vigilax*, with areas that have a greater probability of the occurrence of this species. Thus, the majority of models predict an expanded range for *A. vigilax* in 2085, with some areas showing a higher probabiliy of occurrence,under a high future greenhouse gas concentration scenario.

The economic and human health impact of vector borne diseases underlie the importance of scientifically sound projections of the future spread of common disease vectors such as mosquitos under various climate change scenarios. This is because such information is essential for policy–makers to identify vulnerable communities and to better manage outbreaks and potential epidemics of such diseases. We have shown that using a tool such as the BCCVL can provide the means to robustly and efficiently bracket one source of uncertainty in the future spread of the RRV: the future distribution of a primary mosquito vector of the disease. Research is underway to expand our analysis to include more vector and amplifying host species, as well as future climate projections from a range of different global climate models.

4.   CONCLUSIONS

The BCCVL is a unique tool for the facilitation of research into the impacts of Climate Change on Biodiversity. Strong feedback from researchers in the first few months after the launch of the BCCVL confirms that the goals of lowering the technical requirements for conducting research into climate impacts on biodiversity, as well as reducing the time it takes to do such research, have been met. These two factors are designed to feed into productivity gains for individual researchers, and will likely propel the field forward in terms of the number of species, and species traits which will be the subject of biodiversity-climate change modelling experiments and analyses. As such, we believe that the BCCVL represents a significant step forward for the species distribution and species trait modelling community, and will likely broaden the complexity of the experimental design and the scope of the research undertaken in this field in the future.

The use of the BCCVL to model the current and projected distribution of one of the main species of mosquitos which act as vectors for Ross River Virus, using a large number of algorithms, has effectively illustrated the ability of the BCCVL to bracket the uncertainty surrounding the potential spread of this disease vector in the future, in terms of the choice of algorithm used to model the species, and in terms of the emission scenario used to inform the projections of future climate, upon which the projected distribution is based. Of course, there are multiple sources of uncertainty involved in these predictions- such as climate model or algorithm parameter value choice- both of which will be the focus of future research employing the BCCVL.

The BCCVL is in constant development and current feature limitations of the BCCVL includes the restriction of modelling experiments to the Australian continent, unless users upload their own international environmental datasets. The BCCVL is also constrained by current methods of Species Distribution Modelling, and the issues associated with current SDMs and indeed many environmental models. Future development of the BCCVL will focus on linking to more species occurrence and species trait databases, adding the ability to model multiple species simultaneously within the one experiment, adding more

environmental datasets, and potentially other species distribution models, species trait models, and post-modelling analytical tools.

REFERENCES
Akred J, accessed at https://datascience.berkeley.edu/what-is-big-data) on October 2, 2015.

Gohar L. and Lowe J. (2009) *Summary of the Committee on Climate Change's 2016 peak emission scenarios.* Work stream 1, Report 1 of the AVOID programme (AV/WS1/D1/R01). Available online at www.avoid.uk.net and
http://www.metoffice.gov.uk/media/pdf/5/l/AVOID_WS1_D1_01_20090205.pdf.

Hallgren, W., L. Beaumont, A. Bowness, L. Chambers, E. Graham, H. Holewa, S. Laffan, B. Mackey, H. Nix, J. Price, J. Vanderwal, R. Warren, and G. Weis. The Biodiversity and Climate Change Virtual Laboratory: Where Ecology meets Big Data. In review at Environmental Modelling and Software.

Hilbert DW, Van Den Muyzenberg J (1999) Using an artificial neural network to characterize the relative suitability of environments for forest types in a complex tropical vegetation mosaic. Diversity and Distributions 5:263-274

Laffan, S.W., Lubarsky, E., and Rosauer, D.F. (2010) Biodiverse, a tool for the spatial analysis of biological and related diversity. *Ecography*, 33(4), 643–647.

Lowe, J.A., Huntingford, C., Raper, S.C.B., Jones, C.D., Liddicoat, S.K., and Gohar, L.K. (2009) How difficult is it to recover from dangerous levels of global warming? *Environmental Research Letters*, 4, 014012.

Mahalanobis PC (1936) On the generalized distance in statistics. Proceedings of the National Institute of Sciences (Calcutta) 2:49-55

Mitchell, T. D. (2003) Pattern scaling - An examination of the accuracy of the technique for describing future climates. *Climatic Change* 60, 217-242.

Mitchell, T. D. & Jones, P. D. (2005) An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *International Journal of Climatology,* 25, 693-712.

Nakicenovic, N., Swart, R. (2000) *Special report on emissions scenarios (SRES).* (Working Group III of the Intergovernmental Panel on Climate Change, 2000). Edited by Nebojsa Nakicenovic and Robert Swart, pp. 612. ISBN 0521804930. Cambridge, UK: Cambridge University Press,

Nix, H.A., 1986. A biogeographic analysis of Australian elapid snakes. In: Atlas of Elapid Snakes of Australia. (Ed.) R. Longmore, pp. 4-15. Australian Flora and Fauna Series Number 7. Australian Government Publishing Service: Canberra.

Osborn, T. J. (2009) *A user guide for ClimGen: a flexible tool for generating monthly climate data sets and scenarios.* Climatic Research Unit, University of East Anglia, Norwich.

Peters, D.P.C., Havstad, K.M., Cushing, J., Tweedie, C., Fuentes, O., and Villanueva-Rosales, N. (2014) Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. *Ecosphere*, 5(6), 67.

Phillips, S.J., Miroslav, D., Schapire, R.E. (2004) A Maximum Entropy Approach to Species Distribution Modeling. *Proceedings of the Twenty-First International Conference on Machine Learning*, pp 655-662.

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

Warren, R., de la Nava Santos, S., Arnell, N.W., Bane, M., Barker, T., et al. (2008) Development and illustrative outputs of the Community Integrated Assessment System (CIAS), a multi-institutional modular integrated assessment approach for modelling climate change. *Environmental Modelling Software*, 23, 592-610.

Wigley, T. M. L. & Raper, S. C. B. (2001) Interpretation of high projections for global-mean warming. *Science,* 293, 451-454.

Yu, W., Dale, P., Turner, L., & Tong, S. (2014) Projecting the impact of climate change on the transmission of Ross River virus : methodological challenges and research needs. *Epidemiology and Infection*, 142(10), pp. 2013-2023.