Environmental Modelling & Software 76 (2016) 182–186

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Software data news

The Biodiversity and Climate Change Virtual Laboratory: Where ecology meets big data

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], Gerhard Weis [a]

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

ARTICLE INFO

Article history:
Received 22 May 2015
Received in revised form
24 October 2015
Accepted 29 October 2015
Available online 20 November 2015

Keywords:
Biodiversity
Climate change
Virtual Laboratory
Species distribution modelling

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; and 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, 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.

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". "e-Research" is the application of Information and Communication Technologies (ICT), tools and infrastructure to scientific investigations. In developed countries, e-Research utilises national computing networks, virtual laboratories, research clouds, high performance computing, and "apps" for monitoring and data collection, as well as extensions to citizen science.

Data repositories such as the Atlas of Living Australia (ALA) 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.

Due to the growth in the size, complexity and diversity of datasets ("Big Data"), computational and analytical improvements in statistical and simulation models, such as machine learning (Peters et al., 2014), as well as the recent evolution in web technologies to utilise and work with environmental big data (Vitolo et al., 2015), e-Research has a great potential to advance scientific knowledge, particularly in the field of Ecology. Due to the unprecedented and growing ability to securely store, manage, share,

* Corresponding author. Griffith School of Environment, Gold Coast campus, Griffith University, QLD, 4222, Australia.
E-mail address: w.hallgren@griffith.edu.au (W. Hallgren).

http://dx.doi.org/10.1016/j.envsoft.2015.10.025

analyse and synthesise research data within and across the entire discipline, e-Research has the potential to facilitate research and create new scientific insights in the field of Ecology.

The combination of e-Research and Big Data has made possible the development of the Biodiversity and Climate Change Virtual Laboratory (BCCVL), which we introduce here.

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:

1. A suite of the most commonly used and robust modelling tools and functions to spatially analyse biological data;
2. A comprehensive set of climate change data comprising monthly estimates of downscaled climate change projections of national to international extent;
3. Ancillary physical, environmental, vegetation and land cover data of national extent;
4. Important post-modelling diagnostic, mapping and other visualization capacities;
5. A facility to upload and share data and workflows; and
6. The means to undertake spatial modelling at multiple spatial scales down to a 250 m resolution.

The BCCVL is currently populated with predominantly Australian datasets, but the provision of global datasets, and links to global databases is in development. However, the BCCVL can currently be used for species distribution modelling and other biodiversity analyses for anywhere in the world through the assimilation of user-provided data. Research (in progress) which utilises the BCCVL and user-provided data is focussing on the modelling of international species, thus confirming the BCCVL's utility for conducting biodiversity research internationally. 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 modelling, including data preparation (i.e. identifying, acquiring, scaling/standardising, validating and visualising data), setting up the modelling environment (which could require learning a programming language such as, for example, the R language, then importing data into R, identifying the algorithms to use and the R package for each, importing R package/s into the R environment, running the algorithms (individually), then visualising and manipulating outputs to create maps and graphics), and writing scripts to run complex ensemble experiments.

Depending on the user, this process could take from weeks to months. The BCCVL negates the need for this preparatory work, and for advanced programming and modelling expertise. This results in an increase in research capability and efficiency, and this will likely facilitate the development of additional research trajectories currently not feasible due to both the logistical and computational limitations of individuals and many research groups.

The BCCVL allows comprehensive ensemble modelling experiments (already the norm in the climate modelling community) involving large numbers of species, SDM algorithms, climate model projections, and emissions scenarios. This was once logistically quite challenging, but is now accessible via a web browser. The ability to do this easily will enable more comprehensive comparisons of different SDM algorithms, competing sets of potential explanatory variables, and climate impacts, as simulated by a large range of available climate models, emissions scenarios and projection periods. It will also greatly facilitate comprehensive sensitivity analyses on SDM parameter values, which is an example of a research topic that has received little attention previously, due partly to logistical and computational resource constraints.

As such, the increase in the feasibility of large ensemble modelling experiments, together with education about their importance for scientific rigour and the greater possibility of scientific consilience with it's potential importance for policy development and planning, will encourage users to implement these more complex experimental designs.

The factors mentioned above will contribute to an increase in research productivity (in terms of time saved and scientific output) for species distribution modellers who use the BCCVL, 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. In the sections below we illustrate the use of the BCCVL in the Australian environment.

2. Description of the Biodiversity and Climate Change Virtual Laboratory

A link from the BCCVL homepage at www.bccvl.org.au (Fig. 1) 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:

2.1.1. *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, with a choice of maps available), 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 species traits (e.g. functional traits such as seed mass, rooting depth, root type, fire tolerance, which respond to environmental changes); 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 dataset from 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. Datasets currently accessible within the BCCVL are listed in Table A.1 (Appendix A), and the BCCVL has the capacity to add additional datasets as needed.

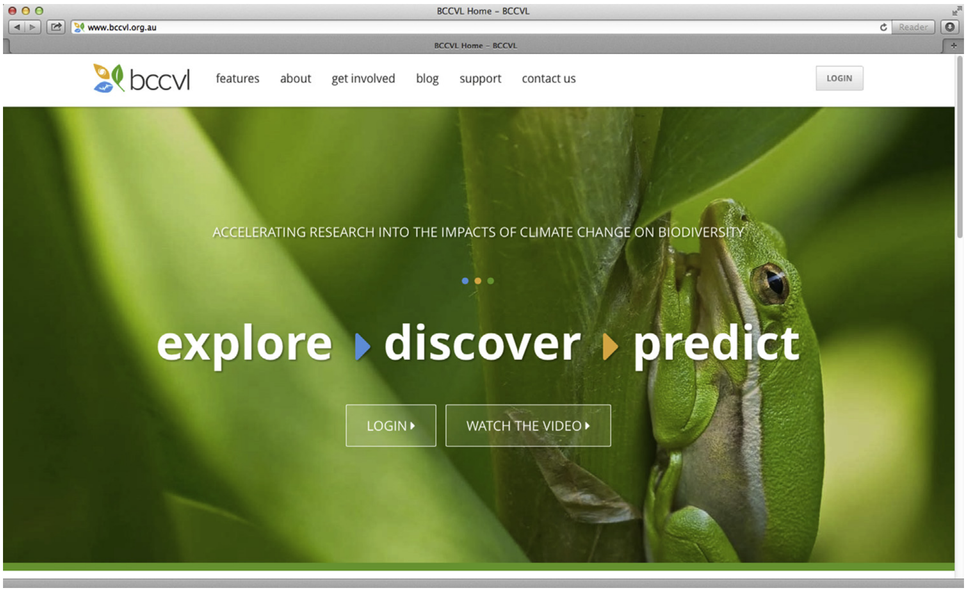

Fig. 1. The BCCVL homepage at www.bccvl.org.au.

2.1.1.1. Climate change projection data. Global climate change modelling was carried out by driving the MAGICC4.1 climate model (Wigley and Raper, 2001; Lowe et al., 2009) with time series of 21st century emissions to create a projection of 21st century climate change. Simulations undertaken explored uncertainties in three key parameters: climate sensitivity, ocean mixing rate, and a climate–carbon cycle feedback factor in MAGICC.

Median projections from these simulations were used to drive a pattern-scaling module ClimGEN (developed from Mitchell, 2003; see also Warren et al., 2008; Osborn, 2009) in which scaled climate change patterns diagnosed from seven of the archived CMIP3 GCM simulations are combined with a baseline climate (CRUTS 3.0 for 1961–1990, updated from Mitchell and Jones, 2005).

The resulting projections of four indicators (monthly mean, minimum and maximum temperatures, defined as the average of the daily maximum (minimum) temperatures during a month, and total precipitation) were downscaled to a resolution of 0.5° × 0.5°. ClimGEN was used to produce projected monthly time series for 30-year periods centred on 2020 (i.e. 2011–2040), 2050, and 2080, These were then averaged to produce representative monthly climates for each 30-year period, using the above mentioned four indicators. This approach was necessary because GCMs have not been run for the mitigation scenarios which were needed for this project.

Emission scenarios used in this analysis included a baseline – SRES A1B (Nakicenovic and Swart, 2000), and several mitigation scenarios developed for the AVOID project (Gohar and Lowe, 2009), which initially follow the baseline scenario before transitioning over seven years so that emissions peak globally in either 2016 or 2030. They are then reduced subsequently at rates of between 2 and 5% annually, until they reach a hypothetical lower limit designed to represent emissions that might be difficult to eliminate, e.g. from the agricultural system. These scenarios are combined with seven alternative GCM-derived change patterns, producing 42 projected climates consistent with the IPCC (PCMDI, 2009).

The downscaled climate data were post-processed to produce eight bioclimatic indices: average maximum temperature of the warmest month, the average minimum temperature of the coldest month, annual mean temperature, temperature seasonality, total annual rainfall, rainfall seasonality, rainfall of the wettest quarter and rainfall of the driest quarter. These were calculated directly from the average climates of the aforementioned 30-year periods. Refer to Appendix A for further details on the methodology.

2.1.2. 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 and endemism) based on species distribution modelling results;
d) *Species Trait Modelling (STM) Experiments* 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.

An example of how to use the BCCVL to implement two of these experiment types; (a) species distribution modelling and (b) projecting a species distribution model into the future with a climate model projection, is given in Appendix B. The SDMs are the core functionality of the BCCVL as their results are used in most of the other components. There are currently 17 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 and Van Den Muyzenberg, 1999), as well as simpler and more easily comprehensible algorithms, such as *Bioclim* (Nix, 1986). All algorithms employed in the BCCVL are described in Table B.1 (Appendix B). A demonstration of a Species Distribution Modelling Experiment is also available in Appendix C.

Supplementary data related to this article can be found online at http://dx.doi.org/10.1016/j.envsoft.2015.10.025.

The algorithms that are currently implemented in species trait modelling within the BCCVL include widely used statistical methodologies such as Generalized Additive Models, Generalized Linear Models (as for species distribution modelling), but 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.

2.1.3. Knowledge base

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, and links to key references and papers. In development are a user-tested, expert-informed 'decision support tool' to guide and inform users as they proceed through the steps to undertake species distribution modelling and other experiments offered by the BCCVL, and an open online course which provides theoretical and practical information on many aspects of the BCCVL experiments. These features 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, which are operated on the Australian National eResearch Collaboration Tools, and Resource Project (NeCTAR) Research Cloud. 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 often designed using commodity hardware (compute, storage and networking) and achieve scale and resilience through the ability to easily add and replace individual components.

The BCCVL utilises the NeCTAR Research Cloud, which will provide 35,000 cores of processing capacity hosted at eight nodes (data centres) distributed across Australia. Utilisation of cloud-based technologies enables the BCCVL to easily scale to meet new demands for processing or storage capacity. For example, the BCCVL uses the SWIFT object storage package to handle and replicate large datasets in a cloud environment. It allows the BCCVL to robustly handle duplication and storage of large datasets with appropriate safeguards to mitigate against data loss and corruption.

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. Appendix D provides a schematic of the major components and information architecture of the BCCVL (Fig. D1), as well as further technical details on the six components that constitute the BCCVL.

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).

3. Conclusions

The BCCVL is a unique tool for the facilitation of research into Biodiversity and the impact of Climate Change. 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 response 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.

Future development of the BCCVL will focus on adding more environmental datasets, and potentially other species distribution models, species trait models, and post-modelling analytical tools.

Acknowledgements

Funding for the BCCVL comes from the Australian National eResearch Tools and Resources Project (NeCTAR). A complete list of participating institutions, development team members, and governance and advisory committee members can be found at www.bccvl.org.au. No co-authors have any conflicts of interests to declare.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.envsoft.2015.10.025.

References

Gohar, L., 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. http://www.metoffice.gov.uk/media/pdf/5/l/AVOID_WS1_D1_01_20090205.pdf.

Hilbert, D.W., 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. Divers. Distributions 5, 263–274.

Laffan, S.W., Lubarsky, E., 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., Gohar, L.K., 2009. How difficult is it to recover from dangerous levels of global warming. Environ. Res. Lett. 4, 014012.

Mitchell, T.D., 2003. Pattern scaling – an examination of the accuracy of the technique for describing future climates. Clim. 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. Int. J. Climatol. 25, 693–712.

Nakicenovic, N., Swart, R., 2000. ISBN 0521804930. In: Nakicenovic, Nebojsa, Swart, Robert (Eds.), Special Report on Emissions Scenarios (SRES). (Working Group III of the Intergovernmental Panel on Climate Change, 2000). Cambridge University Press,, Cambridge, UK, p. 612.

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

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.

PCMDI, 2009. IPCC Model Output. http://www.pcmdi.llnl.gov/ipcc/about_ipcc.php, 26 September 2009.

Peters, D.P.C., Havstad, K.M., Cushing, J., Tweedie, C., Fuentes, O., 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., Anderson, R., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190, 231–259.

Vitolo, C., Elkhatib, Y., Reusser, D., Macleod, C.J.A., Buytaert, W., 2015. Web technologies for environmental big data. Environ. Model. Softw. 63, 185–198.

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. Environ. Model. Softw. 23, 592–610.

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