Sensitivity Analysis to Configuration Option Settings in a Selection of Species Distribution Modelling Algorithms

W. Hallgren [a], F. Santana [b], S. Low-Choy [c], J.H.K. Rehn [b] and B. Mackey [a]

[a] Griffith Climate Change Response Program, Griffith University, QLD
[b] Faculty of Education, Science, Technology & Mathematics, University of Canberra, ACT
[c] Griffith Social and Behavioural Research College, Griffith University, QLD
Email: w.hallgren@griffith.edu.au

Abstract: In pursuit of a more robust provenance in the field of species distribution modelling, an extensive literature search was undertaken to find the typical default values, and the range of values, for configuration settings of a number of the most commonly used statistical algorithms available for constructing species distribution models (SDM), as implemented in the R script packages (such as Dismo and Biomod2) or other species distribution modelling programs like Maxent. We found that documentation of SDM algorithm configuration option settings in the SDM literature is very uncommon, and the justifications for these settings were minimal, when present. Such settings were often the R default values, or were the result of trial and error. This is potentially concerning for a number of reasons; it detracts from the robustness of the provenance for such SDM studies; a lack of documentation of configuration option settings in a paper prevents the replication of an experiment, which contravenes one of the main tenets of the scientific method. Inappropriate or uninformed configuration option settings are particularly concerning if they represent a poorly understood ecological variable or process, and if the algorithm is sensitive to such settings; this could result in erroneous and/or unrealistic SDMs.

We test the sensitivity of two commonly used SDM algorithms to variation in configuration options settings: Random Forests and Boosted Regression Trees. A process of expert elicitation was used to derive a range of appropriate values with which to test the sensitivity of our algorithms. We chose to use species occurrence records for the Koala (*Phascolartos cinereus*) for our sensitivity tests, since the species has a well known distribution. Results were assessed by comparing the geospatial distribution from each sensitivity test (i.e. altered-settings) SDM for differences compared to the control SDM (i.e. default settings), using geographical information systems (QGIS). In addition, two performance measures were used to compare differences among the altered-setting SDMs to the control. The aim of our study was to be able to draw conclusions as to how reliable reported SDM results may be in light of the sensitivity of their algorithms to certain settings, given the often arbitrary nature of such settings, and the lack of awareness of, and/or attendance to this issue in most of the published SDM literature. Our results indicate that all two algorithms tested showed sensitivity to alternate values for some of their settings. Therefore this study has showed that the choice of configuration option settings in Random Forests and Boosted Regression Trees has an impact on the results, and that assigning suitable values for these settings is a relevant consideration and as such should be always published along with the model.

Keywords: *Configuration option settings, provenance, transparency, koala, boosted regression trees*

1. INTRODUCTION

Progress made in computational science in recent years has produced concomitant gains and exciting developments in many scientific disciplines. Many researchers in the computational sciences have increasingly advocated for reproducibility of research, which involves the data and computer code used in a published study to be made available to others, as a minimum standard when assessing the validity of scientific claims. Cassey and Blackburn (2006) have argued that in order for a scientific study to be acceptable for publication, it should be reproducible (NRC, 2003). They contend that "reproducibility is increasingly being requested by journals". This standard of reproducibility is based on the theoretically available detailed log of every action taken using computers, which underlies every computational experiment. If this code were available for scrutiny, such transparency in research would surpass the analogous non-computational experimental descriptions printed in journals using a human language (Peng, 2011).

For scientific fields which inform policy and management decisions, and in which there is difficulty in reproducing field experiments and techniques due to the non-replicable nature of environmental conditions, there is critical need to track the provenance of derived data products and scientific results. 'Provenance', which is "information about entities, activities, and people involved in producing a piece of data or thing, can be used to form assessments about its quality, reliability or trustworthiness" (W3C Working Group, 2013). It should cover initial data collection, quality assurance, analyses, modelling and publication (Reichman et al 2011).

It is with a view towards more robust provenance in the field of species distribution modelling (SDMg), that a literature search was undertaken to find the typical default values, and an appropriate range of values for configuration option settings in each of the statistical algorithms available for constructing species distribution models (SDM) in the Biodiversity and Climate Change Virtual Laboratory (BCCVL) (Hallgren and Mackey, 2014). The BCCVL (www.bccvl.org.au) is an online, cloud-based virtual laboratory which brings together a multitude of datasets, SDM algorithms and several different modelling experiment types to create a highly accessible SDMg platform which can be used to investigate the impact of climate change on species distributions, species traits, and several measures of biodiversity in Australia and around the world (Hallgren et al., 2016). The sources of the default values of packages in R (R Core Team, 2017), which provided the specific implementation of the algorithms used in the BCCVL, were also investigated. Many SDM studies do not publish SDM configuration option settings, and if they do, then justifications for these parameter values, are minimal, if provided, and often rely on the R default values.

It is advisable to document the rationale for SDM configuration option settings for the SDM algorithms implemented in the R script packages (such as Dismo (Hijmans, 2013) and Biomod2 (Thuiller, et al., 2009) or other SDMg programs, like MaxEnt (Phillips et al., 2006). Omitting such documentation detracts from the robustness of the provenance for such SDMg studies. If no information for SDM configuration option settings is provided in a paper, it makes the SDM experiments unable to be replicated (providing a barrier to provenance, and contravening one of the main tenets of the scientific method). Moreover, it works against methodological transparency, which is critical to progress in this field. We contend that it is good scientific practice, and critical for the provenance of a study, to be able to explicitly state and justify the configurable settings used in any modelling exercise.

Configuration option settings which are inappropriate, unrealistic or uninformed, are particularly concerning if they represent a poorly understood ecological variable/process, and also if the resulting SDM is sensitive to their values. Without knowing how sensitive an algorithm is to all configuration option settings, then inappropriately assigned values, i.e. values without reasonable scientific justification, could possibly lead, in an unpredictable manner, to erroneous and/or unrealistic SDM results. The rigour of the modelling process, the validity of the results, and the transparency and provenance of the research may be compromised.

There are a number of studies which investigate the sensitivity of SDM algorithms to one aspect of the modelling process e.g.; sensitivity to pseudoabsence selection in seven different SDM (Barbet-Massin et al., 2012), or sensitivity to input data (Pirathiban et al., 2015), or for one SDM algorithm; (e.g. sensitivity to several model settings for MaxEnt (Merow et al. 2013)). Beaumont et al. (2016) also tested a range of SDMs regarding the 'sensitivity' to projected climate change scenarios in terms of their likelihood to simulate extreme distribution change. However, to date there has been no systematic investigation to sensitivity to a wide range of configuration option settings for many SDM algorithms.

This study reports on a project designed to address this gap. It aims to: (1) Test the sensitivity of algorithms which are widely used in species distribution modelling to variation in their configuration option settings; (2) draw conclusions as to how reliable the resulting SDMs are, in light of this sensitivity, and given the arbitrary

nature of the default settings, and the lack of awareness of, and/or attendance to, this issue in most of the published literature; (3) to articulate why setting reasonable/justifiable settings are important for the algorithms tested - and which configuration options it is most important to be careful setting, given the algorithm's sensitivity to them.

2. METHODOLOGY

Since different statistical models, of vastly different forms, may fit the same process equally well, with each providing a different perspective or "angle" onto the same phenomenon, it is not possible to know ahead of time, for a particular species, which "shape" or template for a SDM will fit well. For this reason, we control for this uncertainty by investigating more than one SDM algorithm for our modelled species.

In this initial study, we examined two machine-learning algorithms for SDM (Franklin, 2010), which are both extensions of Classification Trees (CT): Boosted Regression Trees (BRT: Breiman, 2001) and Random Forests (RF: De'Ath, 2007). Both algorithms are implemented in the BCCVL. A CT defines a sequence of decision rules to define environmental profiles (e.g. low minimum temperatures, under 10°C). Each profile has a different probability of koala presence, defined by different ranges of climate for occurrence or pseudo-absence sites. We assess sensitivity of each algorithm to configuration options in several categories: *robustness*, *complexity, variable importance* and *sampling strategy* for pseudo-absences. CTs are extremely flexible, so that small changes in inputs often result in marked changes to the fitted tree (Hastie et al., 2012). To improve *robustness*, both RF and BRT resample input data, called `bagging', to provide an average prediction across a number ('maximum #trees') of tree models. However, BRT only bags a certain amount of data (the 'bag fraction'), for a random partition of the data (for a given 'random seed').

For RF, the *complexity* of each tree is affected by the number and size of final profiles, respectively the 'maximum #terminal nodes' (maximum number of terminal nodes that trees in the forest can have) and 'terminal node size' (minimum number of observations in terminal nodes). BRT also considers the number of decision rules required to define any profile ('tree complexity' which control whether interactions between predictor variables are fitted). BRT focuses more effort (defined by 'learning rate' (determines the contribution of each tree to the growing model) and a 'tolerance value' (determines when algorithm stops) on decision rules with high uncertainty, effectively boosting those parts of the tree. In addition, within the 'Biomod2' library, *variable importance* for BRT is determined via cross-validation, calculated for each of a 'number of cross-validation' subsets, that may or may not be able to 'stratify' presences (to ensure each subset contains a certain proportion of presences). The Random Forest algorithm was tested for sensitivity to three configurable options in the BCCVL, and the BRT algorithm was tested for sensitivity to seven configurable options in the BCCVL. The control and altered sensitivity test values for these settings are shown in Table 1. Each option needed a different strategy for choosing values, depending on its type. For instance tolerance (a positive real) and maximum number of trees (a natural number) were both better assessed at increasing orders of magnitude, rather than on some linear scale. Some settings were categorical, e.g. whether to stratify prevalences is dichotomous. Some settings manage computational overheads, and are therefore only considered if the model provides poor fit.

We also assessed two configuration options that affect the pseudo-absences generated for any SDM algorithm. The first is the sampling intensity of pseudo-absences generated, in relation to observed occurrences ('absence:presence ratio'). In this study we sought an Australia-wide SDM, to compare with published literature. For this reason, the Surface-Range Envelope (SRE) algorithm (Araujo and Peterson, 2012) was used to generate pseudo-absences from environments dissimilar from occurrences, falling beyond the outer quantile (e.g. 2.5^{th} and 97.5^{th}) for any environmental gradient. The second setting therefore specifies the SRE quantile used (e.g. 0.025 and 0.975): more extreme values (close to 0 or 1) specify absences come from environments that are more dissimilar to presences.

We deliberately chose a relatively simple experimental design for our sensitivity analyses, as a proof of concept. A control model of Koala (*Phascolartos cinereus*) distribution was built using the default configuration option settings of each algorithm. We chose to use species occurrence records (presence points) for the Koala for our sensitivity tests, since the species has been widely studied. We have not used true absence points. All SDM experiments were implemented with ecologically appropriate domain constraints to correspond with the known distribution of Koala (i.e. an ecoregion or bioregion) – this is to constrain the placement of pseudo-absence points generated by the BCCVL. A process of expert elicitation was used to derive a range of appropriate values with which to test the sensitivity of our algorithms (Al-Khairy, 2017).

To evaluate the difference between the control SDMs (i.e. those using default configuration options settings) and altered-settings SDMs, we have assessed the results by comparing the geospatial distribution of the control

SDM and the altered-settings SDM, using geographical information systems (QGIS).

Table 1. Configuration options and test values for (a) both algorithms (affecting the pseudo-absences used as input), (b) Random Forests (RF) and (c) Boosted Regression Trees (BRT).

	(a) Inputs to all algorithms		(b) Random Forest		
	Absence-presence ratio	Pseudo absence SRE quantile	Max no. trees	Terminal node size	Max. no. terminal nodes*
DEFAULT	**1.0**	**none**	**500**	**1**	**none**
TEST VALUES	0.1	0.05	100	5	6
	0.5	0.10	250	10	12
	1.0	0.20	1000	20	30
	2.0		2000	100	60
	10.0				

(c) Boosted Regression Trees								
	Tree complexity	Learning rate	Bag fraction	No. of cross validations	Prevalence stratify	Max no. trees	Tolerance value	Random seed
DEFAULT	**1**	**0.01**	**0.75**	**10**	**Yes**	**10000**	**0.001**	**1**
TEST VALUES	5	0.05	0.9	4	no	1000	0.1	5
	10	0.01	0.95	10		3000	0.05	10
	15	0.005	0.99	20		10000	0.01	20
	20	0.001	0.8	100		30000	0.001	100
		0.0005	0.5				0.005	
			0.2					
			0.1					

* The maximum number of terminal nodes for Random Forests depends on the number of predictor variables used to construct the model, in this case, 6.

We have also noted the differences between the control model and altered-settings BRT models in terms of two evaluation statistics (Table 2). We chose two examples: the false discovery rate (proportion of predicted presences that are observed absences) and misclassification rate (proportion of incorrectly predicted cases), can illustrate how alternative configuration options settings can impact these metrics (Low-Choy, 2015).

Table 2. Evaluation statistics for BRT, for all sensitivity tests, where statistics showed some sensitivity.

Sensitivity Tests	Optimum threshold value:	False Discovery Rate (FDR)	Misclassification Rate
Control	0.336	0.501	0.501
P-A Ratio 0.1	0.585	0.091	0.091
P-A ratio 10	0.216	0.686	0.686
P-A ratio SRE 0.05	0.335	0.501	0.501
Learning Rate 0.0005	0.49	0.501	0.501
Learning Rate 0.05	0.089	0.501	0.501
Trees 1000	0.335	0.501	0.501
Tolerance 0.01	0.335	0.501	0.501
Tolerance 0.05	0.337	0.501	0.501
Tree Complexity 5	0.316	0.501	0.501
Tree Complexity 20	0.312	0.501	0.501

Occurrence datasets for the Koala were downloaded from the Atlas of Living Australia (ALA) and cleaned for duplicates, anomalous occurrence points (such as those from zoos, herbariums, etc., or caused by different geographic coordinate systems), unnecessary replicates in space and time, and for appropriate dates of occurrence. Appropriate habitat predictors (e.g. environment and climate variables) were derived from literature on the climatic conditions that affect the diet composition and physiological stress of koalas (Davis et al., 2014). Survivorship of Koalas is directly and indirectly determined by the frequency, intensity and duration of extreme events such as heatwaves, drought and humidity (Gordon, 1988; Seabrook et al., 2011, Adams, 2010), water availability and rainfall, all of which can affect Koala's distribution, density, habitat preferences, home range sizes, habitat quality, physiological stress (Davies et al., 2013b,c; Sullivan et al., 2003a), and food resource availability (Gordon 1988). Therefore, these predictor variables comprised the WorldClim current climate (Hijmans et al., 2005) bioclimatic variables #5 (Max Temperature of Warmest Month), #9 (Mean Temperature of Driest Quarter), #10 (Mean Temperature of Warmest Quarter), #12 (Annual Precipitation), #14 (Precipitation of Driest Month), and #17 (Precipitation of Driest Quarter). All of these were available in the BCCVL.

3. RESULTS

The Random Forest (RF) algorithm showed sensitivity to different settings for the 'maximum number of terminal nodes' configuration option (from 60 down to 6), with noticeable increases in modelled Koala distribution indicating that predicted Koala distribution increased as tree complexity reduced. RF showed less sensitivity to different test values for the 'terminal node size' configuration option. The area of predicted Koala distribution increased with the 'terminal node size' configuration option (varying from 5 to 100), as a different measure of the tree complexity reduced. RF showed a low level of sensitivity to different test values for the 'maximum number of trees', with perceptible differences in modelled Koala distribution at very small scales with different test values of 100–2000, compared to the default value of 500.

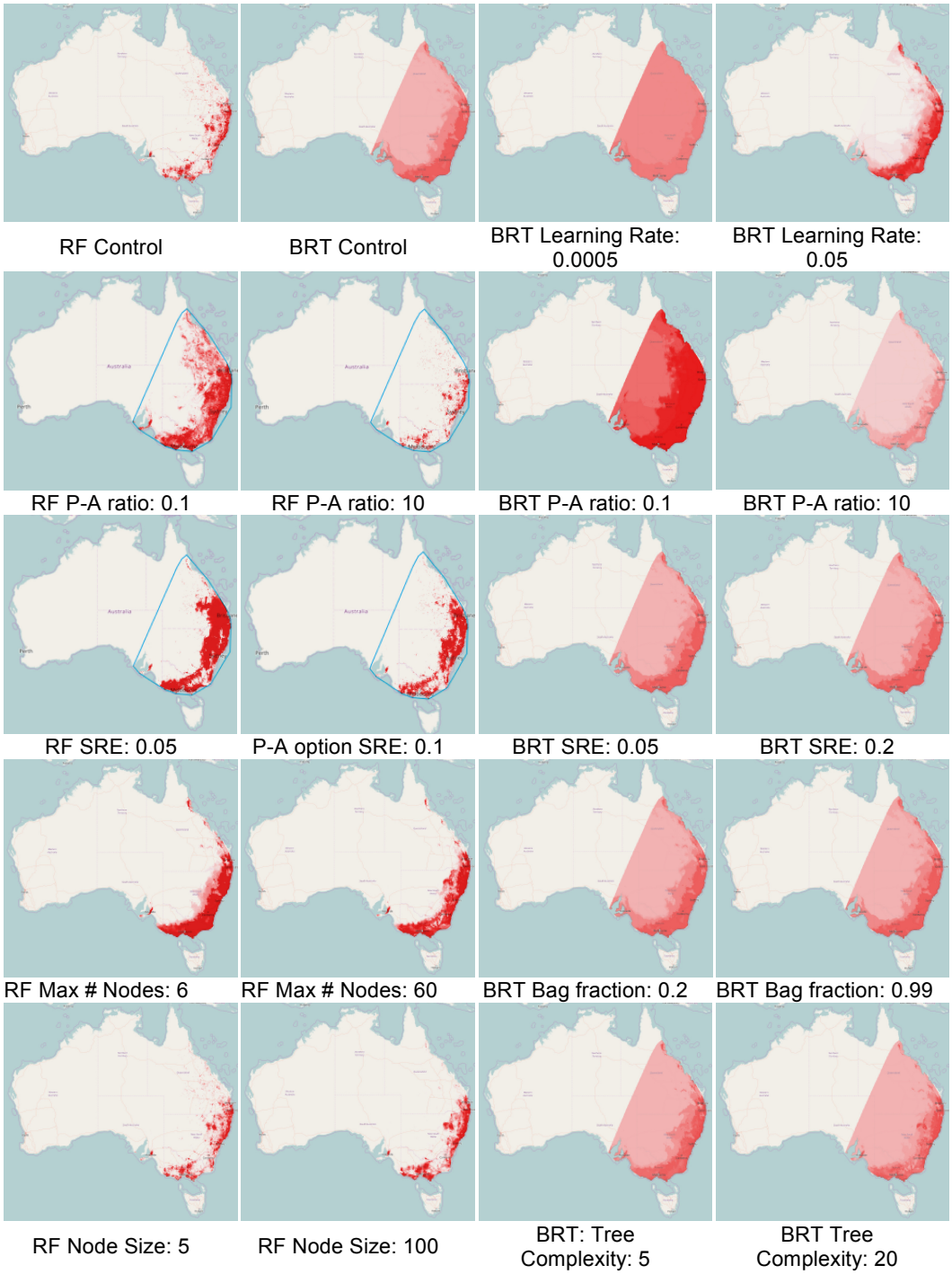

Figure 1. Projected distribution of Koala as a result of using alternative configuration option settings in Random Forests and Boosted Regression Trees.

The BRT algorithm showed sensitivity to different test values for the 'tree complexity' option, with a noticeably expanded Koala distribution with lower values for this option, indicating an inverse relationship, i.e. an increase in area of Koala distribution (red areas in Fig. 1) with decreasing complexity (i.e. from 20 to 5). There was minimal sensitivity shown to the 'learning rate' option for test values from 0.0005 to 0.001, but there was a noticeably expanded Koala distribution with a test value of 0.05, i.e. an increase in probability of Koala distribution along the east and southeast coast of Australia (brighter red areas in Figure 1) with decreasing probability of distribution in inland areas.

The BRT algorithm showed no noticeable sensitivity to robustness options: different test values for the 'maximum number of trees', 'tolerance value' and 'prevalence stratify' options (among those values or options tested), with no apparent change in Koala distribution as values of these options increased from 1000 to 10000 (trees), from 0.005 to 0.1 (tolerance), and from 'yes' to 'no' respectively (prevalence stratify). There was some sensitivity shown to different test values for the 'bag fraction' option, with an almost unnoticeably expanded Koala distribution, particularly on Queensland coast, with (only) the lowest test value of this option, (e.g. bag fraction = 0.1) and then no change in the area of koala distribution from values of bag fraction of 0.2 to 0.9.

The BRT algorithm showed minimal sensitivity to different test values for the 'number of cross validations' option (i.e. that should be created for training and testing the model), with no discernable change in Koala distribution from increasing values of this option setting from 4 to 100. There was noticeable sensitivity to different test values for the 'absence-presence ratio' setting with a marked decrease in Koala distribution (and lower probability of occurrence) from increasing values of this option from 0.1 to 10 (see Fig. 1). Minimal sensitivity was shown to different test values for the 'the quantile for the SRE' setting for the pseudo-absence strategy option, with only small discernable changes in Koala distribution from increasing values from 0.05 to 0.2. The evaluation statistics we chose to use as examples, illustrate that that the alternative configuration option settings do indeed have some impact on these statistics.

4. DISCUSSION AND CONCLUSIONS

An aim of our study was to be able to draw conclusions as to how reliable SDMs are in light of the sensitivity of their algorithms to certain settings, given the often arbitrary nature of such settings, and generally infrequent attendance to this issue in most of the published literature. With respect to an algorithm's configuration, we suggest that SDM studies that have not chosen appropriate settings may yield results that are inaccurate, to the extent that the algorithm is sensitive to those settings. At the very least, this source of uncertainty should be acknowledged and minimized.

Our results indicate that not all configuration options show the same sensitivity, and that of the two algorithms tested, BRT showed sensitivity to alternate settings for five configuration options while RF showed sensitivity to three configuration options.

It is the case for many of the configuration options investigated, particularly for BRT and RF; that the choice of settings for these options can impact the resulting projected distribution markedly, as well as the evaluative statistics, and hence that care must be taken to choose sensible and justifiable values for SDM algorithms.

Another aim of this study was to define which configuration options need setting thoughtfully, given an algorithm's sensitivity to them. Our results suggest that for BRT, the configuration options to be most careful when setting: complexity via the 'maximum number of terminal nodes', 'terminal node size', 'tree complexity; the level of robustness to trees via 'learning rate' and 'bag fraction'; and sampling intensity via 'absence-presence ratio' parameters. For RF, our results indicate that care should be taken when setting the complexity via 'maximum number of terminal nodes, 'terminal node size', and 'maximum number of trees' parameters.

We note that all sensitivities are relative to the control model chosen. Further work will examine broader sensitivity across multiple controls.

ACKNOWLEDGMENTS

The BCCVL is 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.

REFERENCES

Alkhairy, I., Low-Choy, S., Hallgren, W. (2017). Designing elicitation of expert knowledge into conditional probability tables in Bayesian networks: choosing scenarios, submitted to *MODSIM 2017*.

Adams, R.A. (2010). Bat reproduction declines when conditions mimic climate change projections for western North America. *Ecology* 91, 2437–2445.

Araujo M.B., Peterson AT (2012). Uses and misuses of bioclimatic envelope modeling. *Ecology* 93(7): 1527-1539.

Breiman L (2001). Random forests. Machine learning, 45(1): 5-32.

Davies, N., Gramotnev, G., McAlpine, C., Seabrook, L., Baxter, G., Lunney, D., Rhodes, J., Bradley, A. (2013a). Physiological stress of koala populations near the arid edge of their distribution. *PLoS ONE* 11:1–12.

Davies, N., Gramotnev, G., Seabrook, L., Bradley, A., Baxter, G., Rhodes, J., Lunney, D., McAlpine, C. (2013b). Movement patterns of an arboreal marsupial at the edge of its range: a case study of the koala. *Movement Ecol*. 1, 2.

Cassey, P., & Blackburn, T. M. (2006). Reproducibility and repeatability in ecology. *BioScience*, 56(12), 958-959. Available at: http://bioscience.oxfordjournals.org/content/56/12/958.short

De'Ath G (2007). Boosted trees for ecological modeling and prediction. Ecology, 88(1): 243-251.

Gordon, G. (1988). A koala (*Phascolarctos cinereus Goldfuss*) population crash during drought and heatwave conditions in south-western Queensland. Austral. J. Ecol. 13, 451–461.

Hallgren, W.S., Mackey, B. (2014). Analysis of parameter values for the configuration options of statistical Species Distribution Models. T*echnical Report of the Griffith Climate Change Response Program*. Griffith University, Australia, 10pp.

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

Hastie, T., Tibshirani, R., Friedman, J. (2002). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Springer-Verlag. 764 pp.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis A., (2005). Very high resolution interpolated climate surfaces for global land areas. Int. J. of Clim. 25: 1965-1978.

Low-Choy, S., (2015). Getting the Story Straight: Laying the Foundations for Statistical Evaluation of the Performance of Surveillance, chapter 3 in *Biosecurity Surveillance: Quantitative Approaches*, eds. F. Jarrad, S. Low-Choy, K. Mengersen, CAB International, pp 43-74.

Peng, R. D. (2011). Reproducible research in computational science. *Science*, 334(6060), 1226-1227.

Phillips, S. J. et al. (2006). Maximum entropy modeling of species geographic distributions. Ecol. Model. 190: 231–259.

Pirathiban, R., Williams, K. J., Low-Choy, S. J., (2015). Delineating environmental envelopes to improve mapping of species distributions, via a hurdle model with CART &/or MaxEnt, *in* MODSIM 2015 Proceedings, eds. Weber, T., McPhee, M. J., Anderssen, R. S., pp 1455-1461.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science*, 331(6018), 703-705. Available at: http://science.sciencemag.org/content/331/6018/703.full

Seabrook, L., McAlpine, C., Baxter, G., Rhodes, J., Bradley, A., Lunney, D. (2011). Drought-driven change in wildlife distribution and numbers: a case study of koalas in south west Queensland. *Wildlife Res*. 38, 509–524.

Sullivan, B.J., Baxter, G.S., Lisle, A.T., (2003a). Low-density koala (*Phascolarctos cinereus*) populations in the mulgalands of south-west Queensland. III. Broadscale patterns of habitat use. Wildlife Res. 30, 583–591.

W3C Working Group (2013, 30 April) An Overview of the PROV Family of Documents, Available from: https://www.w3.org/TR/prov-overview/PROV-Overview, in W3C Working Group Note, Groth, P., Moreau, L. (Eds), Provenance Working Group, URL public-prov-comments@w3.org