Manaaki Whenua
Landcare Research

Evaluation of online toolboxes for implementing species distribution modelling

Prepared for: Manaaki Whenua – Landcare Research

March 2018

Evaluation of online toolboxes for implementing species distribution modelling

Contract Report: LC3251

James K. McCarthy, Jerry A. Cooper, Susan K. Wiser

Manaaki Whenua – Landcare Research

Reviewed by:

Tom Etherington
Scientist
Manaaki Whenua – Landcare Research

Approved for release by:

Sam Carrick Portfolio Leader – Characterising Land Resources
Manaaki Whenua – Landcare Research

Contents

Summary

Project and client

- Accurately predicting the distribution of species across landscapes at various scales is a fundamental goal in ecology and an active area of research. Species distribution models (SDMs) are quantitative tools that relate the occurrence or abundance data at known locations of individual species (distribution data) to information on the environmental characteristics of those locations.
- SDMs have a wide range of applications, including predicting the impacts of climate change on biodiversity, managing threatened species, predicting the distributional trajectories of invasive species, identifying sites for biological control, and identifying sites appropriate for species reintroduction.
- The SDM process uses computer algorithms to predict a distribution of species in geographical space based on their known point locations across environmental space. They are becoming increasingly important in applied ecology, with new and revised methods under frequent development.
- In this report we first review the background of SDMs and the most commonly applied approaches. We also discuss the types of predictors needed to construct SDMs and the most commonly used species location data. We then evaluate four online turn-key toolboxes for implementing SDMs.
- This report satisfies the requirement of Critical Step 1.2.3 'Evaluation of environments to model and quantify extent of full natural range occupied' for the research programme 'Innovative data analysis for reporting and decision making', funded under MBIE contract to Manaaki Whenua – Landcare Research (PROP-38356-ETR-LCR).

Objective

- To evaluate at least two online toolboxes for implementing SDM modelling that could be integrated into a national information infrastructure.

Methods

- We reviewed the most commonly applied approaches and modelling steps used in SDMs.
- We also reviewed the types of predictors needed to construct SDMs, and the most commonly used types of species location data.
- We then evaluated four online 'turn-key' toolboxes for implementing SDMs, focusing on the feasibility of adapting these toolboxes for New Zealand, incorporating both New Zealand bio- and environmental data. These were:
    - the Atlas of Living Australia
    - the Biodiversity and Climate Change Virtual Laboratory (BCCVL)
    - *Wallace*
    - Lifemapper.

Results

- Four types of species occurrence data can be used to develop SDMs. There are three major algorithms for presence-only models (Convex Hull, BIOCLIM, DOMAIN), seven for presence-background models (MaxEnt, MaxLike, Regression methods, SPP, NPPEN, ENFA, GARP), eight for presence-absence models (GRM, ANN, BRT, Random Forests, MDA, MARS, iCAR, SVMs), and two for occupancy-detection models (false positive occupancy models, O-D).
- The limitations of input species data that must be considered before modelling include (i) the ability to estimate site prevalence, (ii) the impact of imperfect detection, and (iii) the impact of sampling bias.
- A broad range of environmental and biotic drivers of species occurrence could potentially be incorporated as predictors into SDMs, including climate, topography and soil characteristics, indicators of competition, predation, herbivory and mutualisms, disturbance history, and dispersal limitation. Currently there are limitations caused due to lack of data and the ability of SDM algorithms to incorporate such predictors in an appropriate way.
- Turn-key systems differ markedly in the number of different SDM algorithms they support. All systems reviewed here allow data (both species occurrence and predictors) to be either incorporated from existing networks or supplied by the user, with some limitations on the latter. All provide some tools for assessing model fit, but none allow assessments of primary data, such as outlier detection. Only the BCCVL allows ensemble models to be generated and SDM outputs to be readily compared across different analyses. All models generated could be reproduced. The systems could be applied to New Zealand if additional species data and predictor data were supplied.

Conclusions

- Turn-key systems are relatively easy to use, and they produce appealing graphical outputs, extending the accessibility of the approach to many more users. However, they result in serious concessions, because many (or most) of the myriad of adjustable SDM settings are concealed or inaccessible to the user.
- A 2013 review found that the key components of the model building process – such as evaluation of model fit and performance, uncertainty assessment, and inspection of response curves – were not available in many turn-key SDM applications.

Recommendations

- We do not advocate that Manaaki Whenua – Landcare Research (MLWR) should support the development and subsequent use of potentially over-simplified tools, such as 'black box' turn-key modelling software, by end users to support conservation decisions.
- MLWR needs to promote a wider recognition that SDMs should be developed by experts with clear knowledge of the target species and statistical approach. This would be best achieved by continuing to support the development of in-house expertise in

SDMs and ensuring that outputs of SDMs that can support decision-making are widely publicised and made available.

- There is little interest among MWLR staff in using turn-key systems, so we do not recommend tailoring existing turn-key systems for use in New Zealand for internal MWLR purposes.
- A current barrier to the development of credible SDMs in New Zealand is lack of ready access to both species and environmental data.
- MWLR is well placed to:
    - further develop effective data delivery pipelines
    - provide primary data for users to incorporate into their own modelling (e.g. make data available to be harvested by R packages in the same way that the R-package dismo has direct access to GBIF point records)
    - provide appropriate spatial covariate data for users to incorporate into their own modelling with minimal pre-processing (e.g. allow automatic scaling of layers to consistent resolutions and extent).

1 Introduction

Accurately predicting the distributions of species across landscapes at various scales is a fundamental goal in ecology and an active area of research. Species distribution models (SDMs) are quantitative tools relating occurrence or abundance data at known locations of individual species (distribution data) to information on the environmental characteristics of those locations. These types of models are also referred to as 'ecological niche models', 'habitat distribution models', 'resource selection functions', and 'bioclimatic envelopes' (Elith & Graham 2009).

SDMs have a wide range of applications including predicting the impacts of climate change on biodiversity, managing threatened species, managing threatening processes, predicting the distributional trajectories of invasive species, identifying sites for biological control, identifying sites appropriate for species reintroduction, detecting un-surveyed sites with high potential to support rare species, managing interactions between predators and prey species or herbivores and target plants, assessing environmental impacts, and predicting past distributions of organisms, to name a few (Guillera-Arroita et al. 2015).

The SDM process uses computer algorithms to predict a distribution of species in geographical space based on their known point locations across environmental space. The environment is in most cases represented by climate data (such as temperature and precipitation), but other variables such as soil type and land cover can also be used. If models are based on environment alone, they predict locations where the species could occur in the absence of dispersal barriers, historical constraints, or biotic interactions. Incorporating variables that depict these constraints (e.g. disturbance history, current vegetation cover, presence of competing species) allows the full natural range to be predicted.

The use of SDMs is rapidly expanding and new methods are continually being developed, including more recent advances in techniques aimed at predicting multiple species simultaneously while incorporating biotic interactions (joint species distribution models; Warton et al. 2015). Robust and repeatable SDMs will require consensus on and standardisation of methodologies, which does not currently exist despite active debate.

SDMs are becoming increasingly important in applied ecology, with new and revised methods under frequent development. Most of these methods are accessed using computer code/scripts run in command line interfaces (e.g., R and Python) or using more user-friendly software with graphical user interfaces (sometimes referred to as 'turn-key' approaches). Coded scripts are flexible, powerful, and provide high standards of reproducibility and transparency; however, code is often poorly documented and difficult for inexperienced users to run and customise. Turn-key approaches are usually much easier to navigate and use but are usually less customisable. Also, because most cutting-edge approaches are developed using computer code, there is often a lag before the latest approaches are made available with a turn-key interface.

There are also concerns about the lack of reproducibility of turn-key approaches, which are often unable to provide appropriate documentation of the steps taken during analysis (Hampton et al. 2015; Borregaard & Hart 2016). Plus there is the potential for

inexperienced users to misapply methods that are not fully understood. Despite these disadvantages, turn-key approaches make complicated methodology accessible to a much greater audience, and there have been efforts to increase their customisation and reproducibility (e.g. *'Wallace'*; Kass et al. 2018a).

In this report we first review the background of SDMs and the most commonly applied approaches. We also discuss the types of predictors needed to construct SDMs, and the most commonly used species location data. We then evaluate four online turn-key toolboxes for implementing SDMs: Lifemapper, The Atlas of Living Australia, *Wallace*, and The Biodiversity and Climate Change Virtual Laboratory (BCCVL). Our evaluation will consider the feasibility of adapting these toolboxes for New Zealand, incorporating both New Zealand bio- and environmental data.

2 Methods

In this report we review the background of SDMs, including the type of species and environmental data used to parameterise different modelling approaches, the impact of data bias on the interpretation of predictions, and the evaluation of these models. We also evaluate four automated SDM applications, referred to as 'turn-key' approaches: Lifemapper (Stockwell et al. 2006), the Atlas of Living Australia (ALA) (Atlas of Living Australia 2018), The Biodiversity and Climate Change Virtual Laboratory (BCCVL) (Hallgren et al. 2016), and *Wallace* (Kass et al. 2018a).

We will document the effectiveness of these systems to encompass:

1 support for alternative modelling approaches, such as MaxEnt, generalised linear models (GLMs) and generalised additive models (GAMs)

2 the ability to integrate with national and international federated data networks, such as the Global Biodiversity Information Facility (GBIF)

3 the ability to incorporate users' own species occurrence and environmental data

4 the ability to assess model fit

5 the ability to assess the influence of idiosyncrasies in primary data (i.e. outliers) on SDM outputs

6 the ability to make comparisons across multiple SDM models (i.e. with different predictor variables, resolutions or time scales, etc.)

7 the ability to assess the impact of data bias (spatial, temporal) on results

8 reproducibility

9 the spatial extent of applicability (e.g. a specific continent, global, etc.).

To evaluate each turn-key application we ran models across five different terrestrial organisms, covering a range of taxa: a woody plant native to New Zealand and Australia (mānuka, *Leptospermum scoparium*), a globally distributed herb (*Plantago lanceolata*), a common urban bird (house sparrow, *Passer domesticus*), and a New Zealand pest animal species (brushtail possum, *Trichosurus vulpecula*). These species were selected because

they represent a range of growth forms and organism types, and, while we do not present the results of these individual species' distributions, running each turn-key approach multiple times across a range of taxa provided a greater opportunity for assessment. Evaluations were all conducted during March and April 2018.

It should be noted that in this report we do not assess the accuracy of the SDM outputs themselves, but rather the ability of the user to interact with the application and critically evaluate the outputs. For the purposes of this report we assume that each approach produces model summaries and outputs consistent with those recommended for each specific SDM technique.

3 Results

3.1 Species distribution modelling background

Although the concept of a map showing the distribution of a species is easy to grasp, the reality of the process required to produce accurate distribution maps is much more complex. Five major categories of questions arise when considering the types of online toolboxes that would be appropriate for automating, or semi-automating, the process of creating SDMs. These are:

1 **typology of SDMs**: what class of SDMs is appropriate given the available species occurrence data?

2 **data bias:** how will issues such as spatial bias be addressed?

3 **required classes of environmental and biotic predictors for SDMs**: what data are available to create accurate SDMs?

4 **SDM approaches**: which specific modelling approach, or combination of modelling approaches, is appropriate?

5 **model credibility and application**: how will models be evaluated in terms of their predictive accuracy?

These questions are reviewed briefly in the sections below.

3.1.1 Typology of species distribution models

There are four types of species occurrence data typically used in SDMs, each with its own suite of techniques developed for its analysis (Guillera-Arroita et al. 2015). The type of species data available and how these are distributed across environmental space are of fundamental importance when selecting an appropriate SDM technique.

- **Presence-only** models use information about sites where the species was detected without taking into account the environmental conditions in the rest of the landscape.
- **Presence–background** models (sometimes also called 'presence-only' models) estimate habitat preferences by comparing the environmental characteristics at sites where the species has been recorded with a random subset of 'background' (or

'pseudo-absence') locations. Guillera-Arroita et al. (2015) state that it is not possible to tell whether a species is rare and well surveyed or common but under-surveyed from presence records alone. A drawback of these models is that they assume sampling is unbiased − an assumption that is rarely met. The consequence is that habitat suitability may be overestimated for environments that have been sampled more intensively and underestimated for those less surveyed. There is also debate over how background locations can be adequately selected when using these models, in terms of the geographical extent that is considered and the appropriate sample size. Both have been shown to have a notable effect on model outputs (VanDerWal et al. 2009; Stokland et al. 2011). Due to the increasing availability of point records from museum or herbarium specimens, these models are commonly used (Guillera-Arroita et al. 2015).

- **Presence−absence** models estimate the probability of observing a species at a site by comparing the environmental characteristics at sites where the species was detected with those at sites where it was not. The effect of sampling bias is less critical because it does not introduce bias in the estimation. Rather, sampling bias can reduce the precision of estimates for those parts of the environmental space that are under-sampled relative to others.
- **Occupancy−detection** models use detection and non-detection records that have been collected in such a way that the detection (or observation) process can be explicitly modelled. This provides information about the probability of detecting the species given that it is present at a site and how that probability may vary from site to site or visit to visit. This allows models to account for imperfect detection in the estimation of species occupancy probability. The effect of sampling bias is similar to that in presence−absence models. These models are very information rich, but data are often not available because it is generally required that sites be revisited or sampling time quantified.

3.1.2 Data bias

When building an SDM it is important to consider three limitations on the type of species data being used, and how these might affect the interpretation of the SDM's estimates (for more discussion, see Guillera-Arroita et al. 2015):

1 Can site prevalence be estimated?

2 What is the impact of imperfect detection?

3 What is the impact of sampling bias?

Estimation of prevalence

The proportion of sampled sites where a species is present is known as *prevalence*. While not always fully appreciated, the outputs of presence−background SDMs are not an estimate of actual *probabilities* of occurrence, but rather a *relative likelihood* of species occurrence given the model. This is because presence-background data cannot reveal whether a species is rare and well surveyed or common but under-surveyed (Guillera-

Arroita et al. 2015). Presence–absence or occupancy–detection data, and an appropriate SDM technique, are required to distinguish between these two possibilities.

Imperfect detection

Surveys frequently fail to accurately detect all species present at a site, and if this is not accounted for SDMs may end up estimating species observations rather than species occurrences (Gu & Swihart 2004). Imperfect detection affects presence–background and presence–absence data sets and SDM approaches, meaning only the relative likelihood of occurrence can be estimated (but see Guillera-Arroita et al. 2015 for exceptions). When occupancy–detection data are available, estimates of occurrence probabilities can be made.

Sampling bias

Presence–background SDM methods assume that data were randomly sampled from sites selected without bias. However, most of the sources of data for these models (herbarium and museum records, citizen science records) are collected opportunistically and have a spatial bias toward easily accessible and public spaces (Guillera-Arroita et al. 2015). Unfortunately, these locations are often correlated with the environmental covariates commonly included in SDMs and can result in biased estimations of environmental relationships. It is difficult to control for this in presence–background SDMs. Sampling bias has a less critical impact on presence–absence or occupancy–detection models because it does not introduce bias in the estimation (Phillips et al. 2009), but it does reduce the precision of the model.

3.1.3 Required classes of environmental and biotic predictors for SDMs

Key environmental and biotic drivers of species' occurrence must be identified and characterised for SDMs to produce realistic predictions. Because SDMs are typically generated from predictions derived from readily accessible spatial layers, the need to include the key environmental gradients often remains unmet.

For example, while climate often drives plant species distribution at continental scales, topography and soil fertility are critical at smaller scales. In New Zealand the contribution from soils is significant (Wardle 1991; McGlone et al. 2010) because complex topographies and intense weathering rates drive large variation in fertility over small distances. Regional-scale studies show clear relationships between vegetation composition and topographic or age-related soil fertility gradients (Richardson et al. 2004; Jager et al. 2015). These gradients are difficult to model because soils data from native ecosystems are sparse, and estimating local topography from remotely sensed data is difficult. Furthermore, existing soil fertility layers used by modellers (Leathwick et al. 2002) do not reflect the many aspects of soil chemistry most important to indigenous forest and shrubland species (e.g. organic phosphorus, pH). Vegetation communities commonly respond to abrupt changes in environmental variables (Brown 1994; Zimmermann & Kienast 1999) and this problem is widespread globally.

In addition to environment, species distributions also reflect biotic processes such as competition, predation and facilitation, and these can also interact with environment. For example, theory predicts that competition intensifies as soil fertility increases, excluding many species from sites that are otherwise suitable, and additional factors such as rainfall and soil fertility can interact strongly, often influencing species richness patterns (Wright 1992). Species interactions continue to represent a significant challenge for SDMs, but the recent development of joint SDM approaches improves the capacity for these processes to be incorporated (e.g. Ovaskainen et al. 2016).

Species distributions may reflect abrupt limits caused by disturbance events such as earthquakes (Wells et al. 2001), glaciation (Richardson et al. 2004) or fire (Fensham et al. 2003), many of which may be historical. As a result, the historical stability of a site can have a strong influence on the accuracy of many SDM approaches. An organism may be absent from a suitable site due to past events, and this can influence SDMs through non-representative model input data (survey data) and predictions. Geographical and biological limits on dispersal can further affect the ability of populations to colonise new locations or re-establish following disturbance.

Lastly, modern biodiversity databases have an unavoidable bias because they cannot include 'lost occurrences', where species once existed but were removed by humans before any records of these occurrences were made. For example, New Zealand was almost entirely forested before human settlement (McGlone 1989) and now the driest, most fertile regions are largely deforested and many species are now rare or absent. The sensitivity of SDMs to including 'lost occurrences' has rarely been evaluated. Palaeo-ecological distribution data from an appropriate time period (e.g. in New Zealand, 1200–1900 AD, corresponding to before and after deforestation but climate invariant) could be incorporated into SDMs to address this issue. These data could include samples from sediment cores, including ancient DNA, pollen and macrofossils (e.g. leaves, fruits, bark).

A comprehensive approach to SDMs would incorporate competitive interactions, dispersal, demographic rates, and abrupt distributional boundaries, using both climate and soil variables as direct predictors and indirect modulators of competitive interactions and demography. Incorporation of palaeo-ecological data could be appropriate to account for lost occurrences.

The issue, however, is that some of these classes of data currently don't exist at the scales and resolution at which modelling is desired (e.g. soil fertility), and SDM approaches that incorporate these factors are still being developed. For example, classical SDM techniques are commonly criticised for their failure to consider these interactions when making predictions (Ferrier & Guisan 2006). In recent years, however, substantial progress has been made in the field of 'joint species distribution modelling', which simultaneously includes both species- and community-level components (e.g. Warton et al. 2015; Ovaskainen et al. 2016), but these approaches are still in their infancy. Joint SDMs remain an area of active research, and their ability to capture both environmental and community processes make them an important factor in the future development of SDMs.

In this report we restrict our evaluation of SDMs to static distribution models. These models are unable to distinguish between the short- and long-term responses of a species to a stochastically changing environment (e.g. seasonal recolonisation of a floodplain after

flooding). Although factors reflecting disturbance, human influence or successional dynamics can sometimes be included as predictors, this is often achieved with increased difficulty. Alternatives include more mechanistic, dynamic simulation models, but since these approaches require intensive knowledge of the species involved, they are generally restricted to a small subset of well-studied species (Guisan & Zimmermann 2000).

3.1.4 SDM approaches

Many SDM approaches have been developed over the last 20-plus years since the advent of modern modelling and mapping of species distributions (Elith & Leathwick 2009). When fitting SDMs it is recommended that the selection of an appropriate method not depend on statistical considerations and the availability of, or familiarisation with, particular techniques alone. For example, the most appropriate model may depend on the response variable (e.g. percentage cover, individual counts or presence/absence of a species). Some models are designed to reflect the theoretical shape and nature of a species' response to the environment, and there is often a trade-off between optimising accuracy and generality. For SDMs, optimising generality is often achieved by selecting appropriate predictor variables performing robust model selection. Table 1 summarises some common SDM approaches.

Table 1. Some commonly used approaches for SDMs

Name	***Description***	***Reference***
	Presence-only	
BIOCLIM	First developed by Nix (1986), BIOCLIM is often referred to as the first true SDM. It predicts distributions using the minimum and maximum values of environmental variables encountered across a species' measured range. The main drawback of this simple model is the imposed shapes, which can be the cause of a non-justified exclusion or inclusion of a geographical point from the predicted distribution. Also, BIOCLIM models can only accept continuous predictor variables and do not consider interactions.	Nix 1986; Busby 1991; history and use reviewed by Booth et al. 2014
DOMAIN	An early SDM technique, DOMAIN is based on a point-to-point similarity metric between a site of interest and the nearest presence record in environmental space. One drawback is that the metrics used do not account for correlation between covariates. Most suitable when there is a limited number of records available.	Carpenter et al. 1993
	Presence–background	
MaxEnt (Maximum entropy)	This is reportedly the most widely utilised method and software for SDMs, with a relatively user-friendly interface, and it has been shown to out-perform many other methods (Elith et al. 2006). The model minimises the relative entropy between two probability densities (one estimated from the presence data and one from the landscape [pseudo-absences]) defined in the space of supplied covariates (predictor variables). Uses a log-linear model.	Phillips et al. 2006; Phillips & Dudík 2008; Elith et al. 2011; Merow et al. 2013

Name	***Description***	***Reference***
MaxLike	This method is similar to the more widely used MaxEnt, but is capable of estimating absolute occurrence probability (the probability that a species is present in a grid cell). Uses a logit-linear model.	Royle et al. 2012; Merow & Silander 2014
Regression methods (e.g. generalised linear models, generalised additive models)	Regression models such as generalised linear models (GLMs) and generalised additive models (GAMs) are frequently used for SDMs, in part due to their strong foundation and use in modelling ecological relationships. Regression methods have a response (species data) and one or more predictor (usually environmental) variables. Because of their ability to model non-linear relationships using piecewise splines, GAMs are capable of modelling more complicated ecological responses than GLMs. However, less-flexible, non-linear relationships can be specified in GLMs through the inclusion of polynomial terms. These models can also be used when absence data are available (substituting for background data/pseudo-absences).	Various methods are discussed in Elith et al. 2006
Spatial point-process models (SPP; e.g. Poisson point process models)	The development of the SPP approach was motivated by perceived issues with the pseudo-absence approach to methods such as MaxEnt. Since SPP models use point locations in continuous environmental space (rather than at the pixel scale, as in MaxEnt), this approach is well suited to presence-only data commonly associated with museum/herbarium records. SPP models are very closely related to MaxEnt and some regression approaches, the advantage being that they are considered more transparent in terms of assumptions and selection of background points through the use of objectively selected 'quadrature points' rather than more traditional pseudo-absences.	Warton & Shepherd 2010; Renner & Warton 2013
Non Parametric Probabilistic Ecological Niche (NPPEN)	This is derived from a test to compare the ecological niche of two species and is based on a simplification of the Multiple Response Permutation Procedures (MRPP) using the Generalised Mahalanobis distance. MRPP tests whether two groups of observations in a multivariate space are significantly separated. In its adaptation in NPPEN, the test is whether one observation of a taxon belongs to a group of reference observations for that taxon.	Beaugrand et al. 2011
ENFA (Ecological Niche Factor Analysis)	Based on Principal Component Analysis, ENFA summarises multiple environmental variables into a few uncorrelated factors that comprise most of the relevant information. ENFA is quantified through indices of marginality (the direction in which the species' niche differs from the 'global' environment) and specialisation (describes how restricted a species' niche is).	Hirzel et al. 2002
GARP (Genetic Algorithm for Rule Set Production)	GARP is a machine learning approach where a genetic algorithm is implemented to identify associations between occurrences and environment. It uses an iterative process of random rule selection and development to produce a set of rules summarising the species' ecological niche. Rules are selected according to their effectiveness (compared with random), and this process is continued for a set number of iterations (Anderson et al. 2009 ran 2,500 iterations), or until convergence. The final model is a set of if-then statements, which determines whether a species is likely to be present in a particular cell.	Stockwell & Noble 1992; Stockwell & Peters 1999; applied in Anderson et al. 2009

Name	*Description*	*Reference*
	Presence–Absence	
GRM (generalised regression models)	GRM methods are commonly chosen for SDM studies because they can incorporate additive combinations of environmental predictors, and have the flexibility to fit non-linear relationships. Generalised additive models (GAMs) provide the greatest flexibility in terms of fitting complicated non-linear responses. Polynomial transformations can also be applied to generalised linear models (GLMs). GRMs can be extended to account for zero-inflated data sets, structured or nested sampling designs, or spatially auto-correlated data.	GLMs: McCullagh & Nelder 1983; Guisan et al. 2002; Wintle et al. 2005 GAMs: Hastie & Tibshirani 1987; Yee & Mitchell 1991; Guisan et al. 2002
ANN (Artificial Neural Networks)	ANNs are machine learning techniques, first developed as models for the human brain and mostly used for the classification of remotely sensed data (Civco 1993; German & Gahegan 1996). They work by deriving composite variables called 'hidden layers', which are weighted, linear combinations of the predictors to model species occurrences. ANNs are often used in image classification and vegetation mapping (see Linderman et al. 2004), with less SDM application. Although the accuracy of ANNs can approach that of more established techniques, the steep learning curve associated with ANNs prevent their wider application.	Olden et al. 2008; Franklin 2009
Boosted Regression Trees (BRT)	Decision trees are machine learning techniques that work by dividing data into subgroups using predictor variables, resulting in a decision tree of binary decision rules that can be used to classify observations and make predictions. 'Boosting' is a form of model averaging that works by repeatedly sampling the data to develop trees using a withheld sample to evaluate the model. 'Problem' observations (those frequently misclassified by previous models) have a higher probability of being selected, aiding in the refinement of the model. Prediction is based on an average of all trees built in the model.	Elith et al. 2008; Franklin 2009
Random Forests	Random Forests is similar to Boosted Regression Trees. It works by building and averaging many (typically 500–2,000) uncorrelated trees, but each split is developed with a random subset of the predictor variables. This method avoids overfitting the data, but due to the large number of trees, results can be difficult to interpret.	Prasad et al. 2006; Franklin 2009
MDA (Mixture Discriminant Analysis)	MDA is a classification SDM method based on mixed models. It assumes that the class of each environmental variable follows a normal distribution, and that each class can be split and modelled as mixtures of subclasses (each also with a normal distribution). The mixture of these normal distributions is used to generate a density estimation of each class, indicating a species presence or absence while accounting for multi-modal distributions.	Hastie et al. 1994; Hastie & Tibshirani 1996

Name	*Description*	*Reference*
MARS (Multivariate Adaptive Regression Splines) & MARS-INT (Multivariate Adaptive Regression Splines with Interactions)	MARS is a non-linear regression method, somewhat related to GAMs because it uses piecewise splines (Leathwick et al. 2006; Franklin 2009). This technique combines the strengths of regression trees and spline fitting by replacing the decision steps of regression trees with piecewise linear basis functions. MARS can model complex relationships between multiple predictors and a response variable in a computationally efficient manner (unlike GAMs, which can be slow with large data sets). Interactions among predictions (MARS-INT) are also considered between the sub-regions of each basis function, rather than globally (as in most models).	Hastie et al. 2001; Muñoz & Felicísimo 2004
iCAR (Bayesian intrinsic conditional autoregressive model)	Intrinsic conditional autoregressive (iCAR) models are network-based models designed to model spatially auto-correlated data based on neighbourhood relationships. These models make use of a spatial weights matrix to quantify the relative effects the spatial dependencies have on the data.	Besag 1974; for a recent discussion see Ver Hoef et al. 2018
SVMs (Support Vector Machines)	Recently developed and applied to SDM, SVMs are typically designed for two-class problems where a hyperplane separates two classes (such as species presences and absences) in predictor space. One-class SVMs can also be effective for presence-only species observations (Drake et al. 2006).	Guo et al. 2005; Franklin 2009
Occupancy–detection		
False positive occupancy models	These models incorporate information about the degree of certainty in detection and allow data from multiple survey/detection methods to be modelled together.	Miller et al. 2011
O-D (occupancy detection)	O-D models are an extension of logistic regression that account for imperfect detection. They are an SDM that separates occupancy, a biological process, from detection, an observational process, to account for false negatives.	MacKenzie et al. 2002

The modelling process itself consists of two important stages. *Model calibration* includes deciding which set of explanatory variables should be included in the model, transforming explanatory variables as required, parameter estimation, measuring model fit, and selecting the best model given the data. Selecting the best model may be based on comparing the fit of competing models, evaluating the influence of outliers and leverage points and, for some model types, using processes such as pruning and cross-validation. It is also recommended that consideration be given to detecting and correcting for potential sampling bias in species record data, especially when using presence-only data (El-Gabbas & Dormann 2018).

The second stage is *prediction* of the potential distribution of the taxon within the modelled area. This stage is typically implemented in a GIS environment, but can also be generated in programs such as R (R Core Team 2016) using spatially complete spatial layers of the explanatory variables used in the model. Although rarely done, mapping of the uncertainty of predictions is also useful (Rocchini et al. 2011; Stoklosa et al. 2015).

While SDMs were originally used for predicting within sampled regions, predictions into new times and locations is becoming a frequent application. SDMs are poorly suited to these applications due to their reliance on correlative statistics with little

mechanistic/causal relationships, but researchers persist with their application into unsampled environments and climates due to a lack of viable alternatives (Elith & Franklin 2013). Common applications include research into the impacts of climate change and invasion, and prediction into past climates.

3.1.5 Model evaluation

Model evaluation is the process of measuring the adequacy of model predictions compared with field observations. Other terms commonly used to describe this step are 'validation' and 'accuracy assessment'. Prediction errors in SDMs can result from model specification errors (incorrect predictor variables or parameters) or data errors (missing predictors). Data errors are usually driven by an incomplete knowledge of the environmental factors that drive a species' distribution, or lack of suitable spatial data that represent those factors.

While most statistical models used to test hypotheses use R^2 and p-values to evaluate model fit, typically these are not appropriate for SDM methods. Rather, modellers usually employ a range of cross-validation techniques using data resampling (Elith & Leathwick 2009). Commonly used approaches are cross-validation and Jack-knife or bootstrapping, but there is little consensus on which is best. These approaches typically involve creating a model with a 'training' data set and testing its predictive accuracy across a withheld 'evaluation' data set. This is usually repeated across multiple iterations. Data sets from different, independent sources can be used for calibration and evaluation, but this is uncommon because it is usually desirable to use as many observations as possible during the calibration stage of the modelling process (but see Franklin 2002; Elith et al. 2006).

Model evaluation can be measured using threshold-dependent or threshold-independent measures of accuracy. Threshold-dependent measures, such as sensitivity, specificity and kappa, are useful for categorical SDM predictions, where a species is predicted as a binary present or not present. However, many SDMs incorporate continuous probability maps of species occurrence, from which a threshold can be applied to produce binary maps. Threshold-independent measures of accuracy have recently become popular, not only to evaluate SDMs but also to compare among SDM methods, candidate predictors, etc. Frequently used threshold-independent measures of accuracy include area under the receiver operator curve (AUC) and correlation coefficients.

3.1.6 Model credibility and application

Judging an SDM's credibility is a subjective exercise and relies heavily on the intended application of the model (Guisan & Zimmermann 2000). For example, if the prime focus is to explore species–environment relationships, a model may be judged as 'good' when its predictions closely match the observed data. However, many studies require accurate predictions of species ranges at high resolutions. Credibility can be based on test statistics (i.e. Kappa; Monserud & Leemans 1992) and should represent a level that is acceptable to the intended user of the output.

SDM outputs are inherently probabilistic, and this needs to be considered in their application (Guisan & Zimmermann 2000). Given the correlative nature of SDMs, extreme

care must be taken when extrapolating into environmental space that has not been sampled; i.e. novel combinations of predictors that did not occur in the data from which the model was derived. Such extrapolation commonly occurs during spatial prediction across regions, or when making predictions into future climates.

While process-based, mechanistic (or semi-mechanistic) models (e.g. Kearney & Porter 2009) may be considered preferable under these circumstances, these are usually only applied to well-studied species and, in the case of forests, to idealised stands restricted to a single species, and they require detailed, often species-specific measurements of physiological function. Unfortunately, these data are rarely available, and indirect, correlative approaches, such as SDM, remain one of the few feasible approaches for extrapolating distributions across time and space. These limitations must be considered when assessing SDM outputs.

3.2 Assessment of turn-key SDM approaches

A summary of the results from our evaluation of turn-key SDM applications is presented in Table 2.

Table 2. Comparison of four semi-automated turn-key species distribution modelling applications

	Atlas of Living Australia (ALA)	***Biodiversity and Climate Change Virtual Laboratory (BCCVL)***	***Wallace***	***Lifemapper***[1]
1. SDM approaches	MaxEnt	17 different algorithms	BIOCLIM and MaxEnt	ANN, BIOCLIM, GARP, SVMs
2. Integration with data networks	Multiple Australian and NZ custodians (universities, CRIs, herbariums, CSIRIO)	ALA, GBIF and Aekos (Australian vegetation plot data set)	GBIF, VertNet (vertebrate occurrence records) and BISON (North American occurrence data)	GBIF
3. Allows incorporation of user's data	Can import species records but not environmental correlates	Yes: both species occurrence data and spatial layers can be uploaded.	Yes: both species occurrence data and spatial layers can be uploaded.	Unable to assess.
4. Ability to assess model fit	Export includes Jack-knife for assessment of environmental covariate importance / contribution to model, and also AUC values/plots.	Depends on the SDM algorithm, but AUC values/plots, variable importance comparisons, and various error rate plots are available for most models.	AUC values/plots are produced for MaxEnt. At present there is no ability to assess the importance of different covariates.	Unable to assess.
5. Assessment of outliers, etc., in primary data	Not formally, but partial dependence plots show responses of species to environmental variables.	Not formally.	Not formally.	Unable to assess.
6. Ability to compare across SDM outputs	Not formally, but multiple outputs can be viewed simultaneously in the online interface and compared visually.	SDM maps and outputs, including response curves, can be plotted side by side (Figure 2b) and ensembles can be generated (Figure 2d).	Not currently.	Unable to assess.
7. Ability to assess impact of data bias	Nil	Nil	Not currently.	Unable to assess.
8. Reproducibility	All model outputs are provided in a ZIP folder, including MaxEnt parameters, a shapefile and partial dependence plots. No R code (or similar) for publication.	Model outputs and parameters are stored in the user's online account and can be exported, along with the R code used to generate the BCCVL results.	An annotated and executable R Markdown file is produced by *Wallace*, enabling the user to re-run analyses. All outputs must be downloaded because there is no cloud storage for *Wallace*	Unable to assess.
9. Spatial extent	Heavily focussed on Australia, with most spatial layers only available for this region (climate layers available globally). Some point records available for other regions, including NZ.	Through access to GBIF, the BCCVL effectively has a global extent. Most spatial layers, excluding climate, are restricted to Australia but additional point records and spatial data can be uploaded by the user.	Global.	Global.

[1]Lifemapper was not functional during testing.

3.2.1 The Atlas of Living Australia (ALA)

The ALA is the Australian node of the Global Biodiversity Infrastructure Facility (GBIF) and is Australia's national biodiversity database, providing free access to millions of occurrence records and hundreds of environmental layers. The ALA has a user-friendly, stable spatial interface (Figure 1a), in which point records can be displayed for multiple species and SDMs can be performed.

Unfortunately, although the ALA contains records of species occurrence outside Australia (e.g. it contains records of New Zealand plant species mobilised by the Australasian Virtual Herbarium network), most environmental layers are only available for Australian boundaries, and users cannot manually upload their own, so, depending on the layers selected, predictions are often restricted to Australia. Layers with wider coverage include the WorldClim series of climate grids (Hijmans et al. 2005), enabling SDMs to be run outside Australia (including New Zealand), but these will be based solely on temperature and precipitation, with no consideration of solar radiation or edaphic variables. While disappointing, this is not surprising for an Australian-developed and -funded database.

The ALA includes only one SDM method, MaxEnt, which can be run directly from the spatial portal (Figure 1a). While there are many parameters that can be adjusted in MaxEnt (see Elith et al. 2011 for a detailed guide to MaxEnt), these cannot be adjusted within the ALA interface, with defaults used for most parameters. The aspects that can be controlled by the user within the ALA are limited and include the following.

1 **The spatial area of interest.** When running an SDM there are several different ways the user can define the area of interest within which the occurrence points will be considered and the species' range will be predicted. This is called the 'Active Area' and can be defined as the displayed map window extent, the Australian region, global, or within a user-defined area of interest. This can be drawn manually as a polygon directly in the spatial portal, uploaded as a shape file, defined as a radius surrounding a point or address, or selected from a mapped polygonal layer.

2 **Included species point records.** The records contained in the ALA have been obtained from a range of sources and there can be errors relating to taxonomy and coordinate uncertainty, and some records may be out of date. Species records used as inputs for SDMs in the ALA can be filtered by date or coordinate uncertainty, though this requires exporting the ALA data, manual filtering and re-importing. Points can also be filtered by location using a polygon that can be defined directly in the spatial portal (or uploaded as a shape file), and further filtered automatically during the SDM fitting process by excluding 'spatially suspect' records (records that fail the ALA's spatial tests; i.e. terrestrial species in the ocean, coordinates given as 0, 0).

3 **Included environmental layers.** The ALA includes around 300 environmental layers that can be included in SDMs, though most of these are restricted to Australia. While selecting environmental layers to use as covariates in the SDM, they can be viewed directly in the portal and a 'traffic light' colour system is used to indicate the degree of correlation between selected variables.

4 **Select SDM verification options.** Before running the SDM, there are three optional MaxEnt parameters that can be selected by the user to assess the fit of the model. A

Jack-knife can be included that examines the significance of the environmental predictors individually. The user has the option to create response curves, including a chart for each layer, which can be used to evaluate how well the model captures the response of the species to that variable (i.e. Figure 1c). Finally, the user can define the percentage of occurrences that are withheld to test the model.

Once the model specifications are selected, the SDM is run on ALA servers and the results are automatically downloaded in a zip file and displayed as a map in the ALA spatial portal. The zip file includes metadata for the model, including the MaxEnt parameters used, model verification results (including AUC statistics), a thumbnail prediction map (Figure 1b) and shape file, covariate response curves (Figure 1c), and an analysis of covariate contributions (Jack-knife results). Also included is a list of species occurrence locations included in the model and a unique identifier that can be used to restore the model, even in a later session. A scatterplot between an SDM-generated layer and an environmental covariate included in the SDM can be generated to visualise the relationship between that covariate and the predicted probability of occurrence at each of the locations where the species was observed (Figure 1d). Points on the scatterplot can also be selected directly to identify the spatial location of outliers and points of interest in the spatial portal.

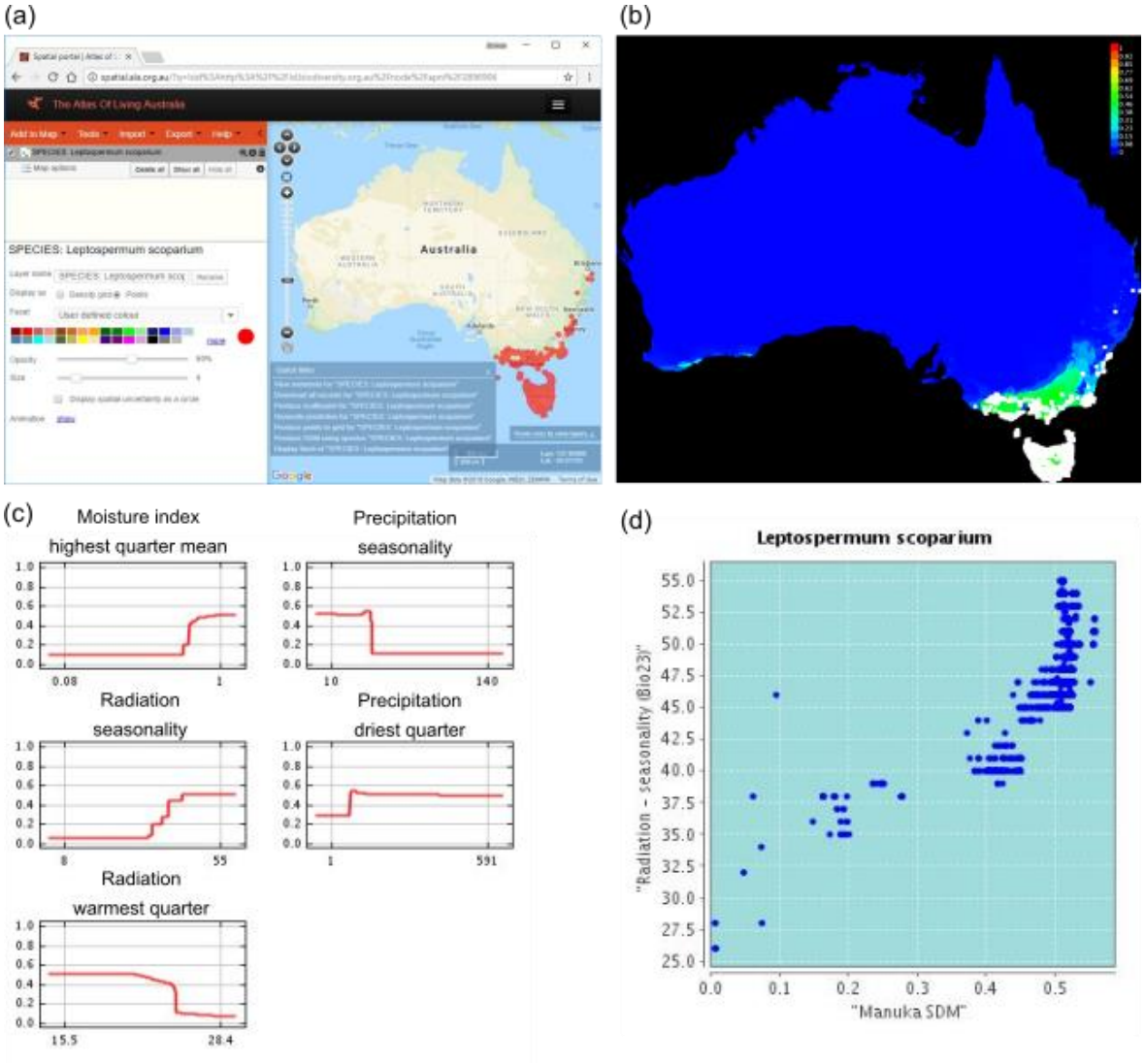

Figure 1. The Atlas of Living Australia (ALA) is the Australian node of GBIF and contains millions of point records and hundreds of spatial layers that can be used for SDM.
(a) The ALA program has a 'spatial portal' where point records can be displayed and SDMs are processed. (b) An example SDM output from an ALA prediction showing the predicted distribution of mānuka (Leptospermum scoparium) across Australia using MaxEnt. (c) Fitted functions of the five environmental variables included in the model. (d) The ALA also has a scatterplot function, which can be used to display values of an environmental variable and the probability of occurrence generated from an SDM output to visualise a species' response.

3.2.2 The Biodiversity and Climate Change Virtual Laboratory (BCCVL)

The BCCVL is a cloud-based research facility developed in Australia with the goal of assisting researchers to access data and high-performance computational resources through a user-friendly graphical user interface that accesses the statistical program R (Hallgren et al. 2016). The BCCVL provides access to species point occurrence records from the ALA, GBIF and Aekos (one of Australia's vegetation plot databases) and contains several vegetation, climate, soil and geology layers. Access to species point records held by ALA and GBIF mean that records are available at a global scale, but most environmental layers are restricted to Australia. Climate variables from the WorldClim series (Hijmans et al. 2005) are available for analyses outside Australia, and users can also upload their own climate and environmental

layers. There are options to automatically scale layers to consistent resolutions while parameterising the SDM.

At present the BCCVL has the functionality to run 17 different SDM algorithms, including MaxEnt, Artificial Neural Networks, and several machine learning and regression techniques including GLMs, GAMs, Random Forests and Boosted Regression Trees (see Hallgren et al. 2016 for a full list). There is also a comprehensive set of training videos available online covering the conceptual background of SDMs, and a tutorial explaining how they can be run in the BCCVL. As with the ALA, the user navigates through a set of menus where input data (species and environmental, including user-uploaded) are selected to be included in the model(s) (Figure 2a). The user also has the option to include true absence data; otherwise background, or pseudo-absence, data are automatically generated with user-specified constraints (i.e. presence−absence ratio, location selection strategy, etc.).

There is a range of methods available to restrict the geographical range of the analyses, including convex hull of occurrence points, or using polygons either defined/drawn manually directly in the BCCVL visualiser or uploaded as a shape file. The user has the option of selecting any number/combination of the 17 SDM algorithms, and these are run concurrently. Each SDM algorithm is configurable: for example, with MaxEnt the user can select the number of background points and, unlike in the ALA, which also uses this method, can select the predictor variable features permitted for the model. Defaults are automatically selected when values are not set by the user.

Once the models are parameterised, the SDMs are run on a cloud-based Australian supercomputing facility called the Australian National eResearch Collaboration Tools and Research Project (NeCTAR). Since the models are run on an external cloud-based network, the user can log off their personal machine after initiation with the outputs saved to the user's BCCVL profile/login. The outputs and model verification statistics from the selected SDM algorithms can be compared within the BCCVL interface (Figure 2b,c) and the R code used to generate results can be exported. Additional features of BCCVL include the generation of an ensemble SDM from multiple algorithms (Figure 2d) and future projections of distribution under climate change.

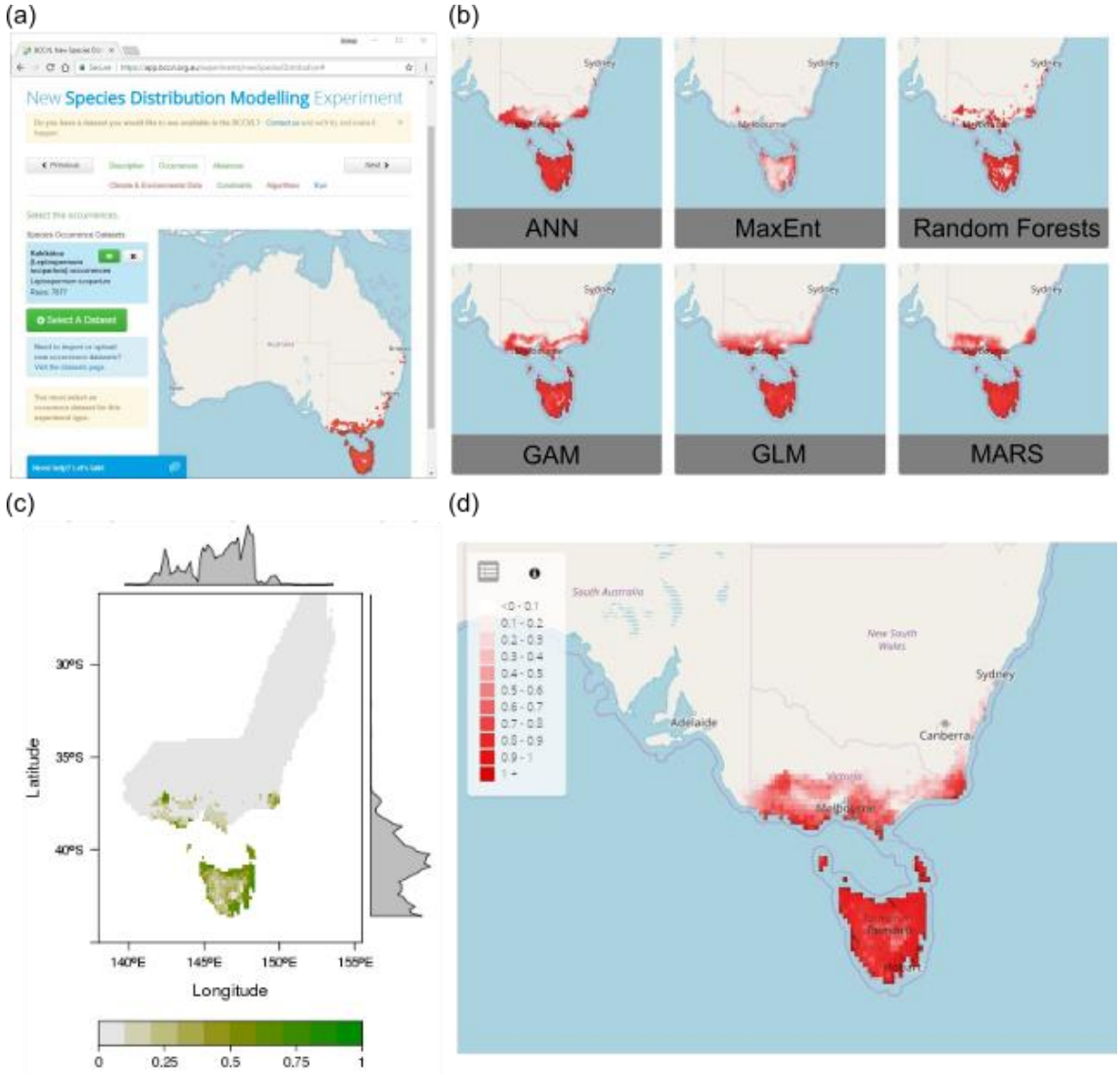

Figure 2. The BCCVL provides access to a large collection of species occurrence records and spatial layers, as well as 17 different SDM tools.
All panels of this figure show point records and predictions for mānuka (*Leptospermum scoparium*). (a) Species observations can be viewed directly within the online BCCVL interface/visualiser (in this case, mānuka, *Leptospermum scoparium*). (b) After running a set of SDM analyses, outputs from different algorithms can be viewed side by side in the BCCVL visualiser for comparison. (c) The user can view projection plots showing densities of occurrence across latitude and longitude. (d) Ensembles can be generated to provide a consensus among multiple SDM algorithms; here we show the mean predicted probability of occurrence for the six algorithms shown in (b), but additional measures (minimum, maximum, variance, various percentiles) can also be generated.

3.2.3 Wallace

Wallace is a recently developed, open-source, R-based platform for reproducible modelling of species distributions (Kass et al. 2018a). It was developed to combine the positive attributes of traditional, command-line interfaces (such as R), which are customisable but complex, and graphical user interfaces, which are usually less flexible. The package is available on both CRAN (Kass et al. 2018b) and Github, and uses a graphical user interface, written using the web app development R package shiny (Chang et al. 2017). *Wallace*

provides access to species point records from GBIF, VertNet (vertebrate occurrence records) and BISON (Biodiversity Information Serving Our Nation: North American occurrence data for species of most taxonomic groups), and spatial data from the WorldClim series at varying resolutions (Hijmans et al. 2005). No habitat, soil and geology layers are provided, but these may be supplied by the user.

Unlike the other turn-key approaches evaluated here, *Wallace* runs directly on the user's PC rather than on a cloud-based server. This is an advantage if the user is running SDMs using spatial data saved on their local PC because it avoids uploading large spatial files over an internet connection. However, this is a disadvantage when using *Wallace*-provided spatial layers because they must be first downloaded to the user's PC (*Wallace* facilitates this but it can take time depending on the speed of the internet connection).

At present, *Wallace* includes two SDM approaches, BIOCLIM and MaxEnt, but the modular nature of the program will allow it to be expanded on and contributed to by the community (Kass et al. 2018a). Unlike BCCVL, different SDM approaches cannot be run concurrently and there is no functionality to run any ensemble analyses. To initiate the *Wallace* graphical user interface, the user must run two lines of code in R (following installation of the package): (1) '`library(wallace)`' to load the package, and (2) '`run_wallace()`' to launch the application in the user's default web browser. While the user interacts with *Wallace* in a web browser, the data processing is completed by a 'background' R-session on the user's PC.

The graphical interface is intuitive and straightforward (Figure 3a), with a similar layout to BCCVL. It is split into a series of nine components, each with one or more modules. All modules, associated R packages and authors are documented in the interface, and referenced information is included to provide the user with background information and troubleshooting tips. The nine components are as follows (Kass et al. 2018a).

1 **Obtain occurrence data.** The user can query a database or upload their own data. Unfortunately, when querying a database the user can only extract a maximum of 500 occurrences, and users can only query one of the three databases at a time. Records are first filtered to remove those duplicates and those without coordinates and then plotted on a map.

2 **Process occurrence data.** The user can choose which point records to include in their analysis by removing occurrences outside a user-defined polygon on a map, by ID, or by using a spatial thin where occurrences are systematically removed to decrease their density.

3 **Obtain environmental data.** *Wallace* provides access to WorldClim climate layers at various resolutions, and also allows the user to input layers.

4 **Process Environmental data.** The user specifies a background extent (study area) to clip predictor layers and draw background samples (if required). The user can upload a shape file, or choose a background extent using a bounding box (Figure 3a), minimum convex polygon or point buffers from the occurrence data. A user-defined number of background samples can also be generated at this stage.

5 **Partition occurrence data.** The user can choose how to partition occurrence points to evaluate the model. Points are sorted, spatially or non-spatially, into a user-selected number of groups used for *k*-fold cross-validation.

6 **Build and evaluate niche model.** The user can select from either BIOCLIM or MaxEnt. For MaxEnt, feature classes can be selected to define the flexibility of the model response and the regularisation multiplier can be selected to penalise complexity. The user does not have the capacity to alter any other MaxEnt parameters.

7 **Visualise model results.** The user can map spatial predictions generated with the SDM (Figure 3b), model evaluation plots and response curves (Figure 3c). These can also be downloaded to the user's PC.

8 **Project model.** The user can project models under alternative climate scenarios, or across new areas.

9 **Session code.** One of the main advantages of *Wallace* is its reproducibility. The user can download a fully executable R Markdown file that can be used to rerun analyses, share results, or provide supplementary information for a research output.

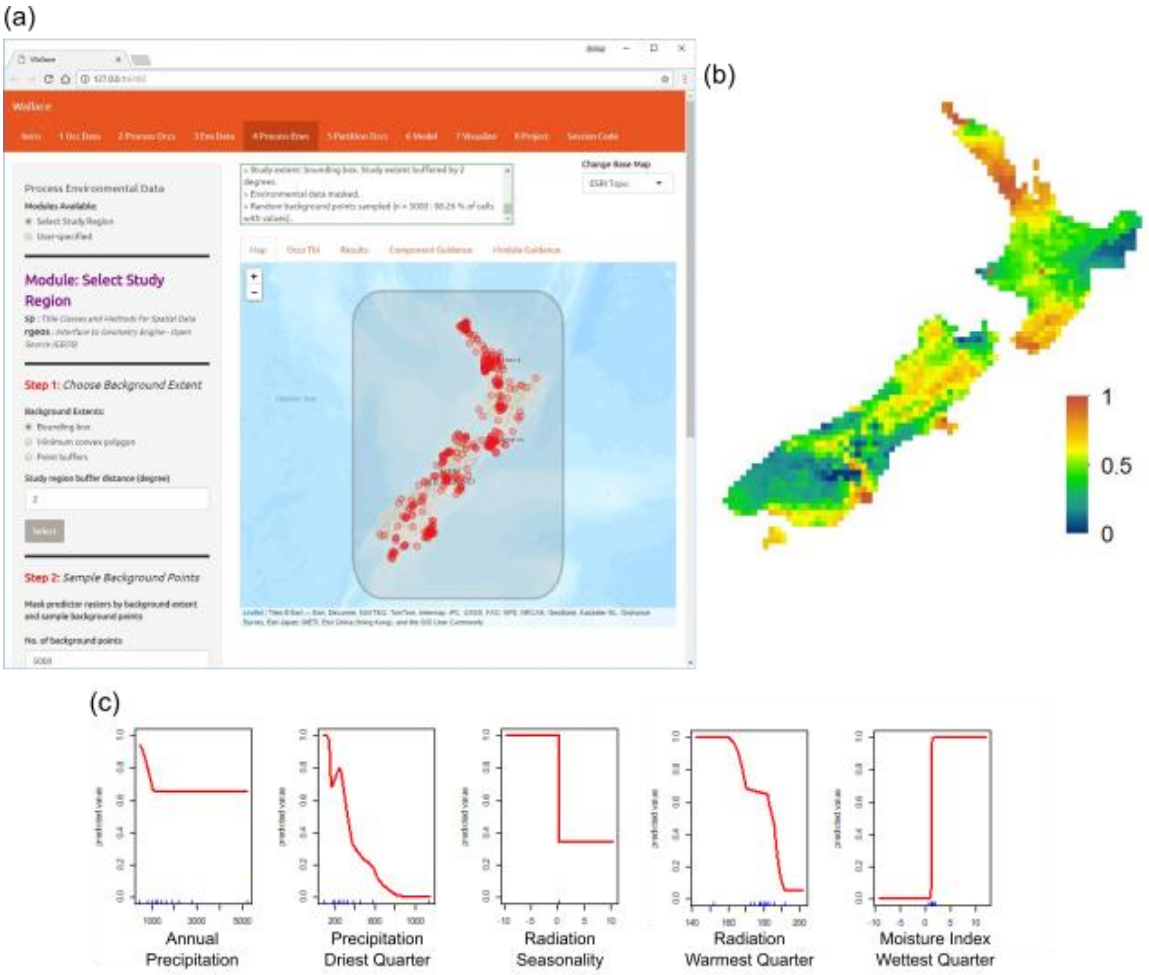

Figure 3. *Wallace* is an open source R package for generating SDMs through a graphical user interface.
All panels in this figure show point records and predictions for mānuka (*Leptospermum scoparium*). (a) The *Wallace* interface is triggered in a web browser after executing two lines of code in the program R (following the installation of relevant packages). All analyses run on the user's PC using their own data, or data available within the program. Displayed here are point records for mānuka extracted from GBIF and the user-defined study region (in grey). (b) MaxEnt output – while output maps from SDMs can be visualised directly within *Wallace*, higher-quality maps can be produced using GIS software. (c) Fitted functions of the five environmental variables included in the model.

3.2.4 Lifemapper

Unfortunately Lifemapper was not operational at the time of evaluation. While the website could be accessed, their modelling component was not operational. All information in Table 2 was sourced from parts of their website that were operational.

4 Discussion

Following substantial developments in analytical techniques and computing power, the complexity of ecological analyses has greatly increased in recent times (Bolker 2008; Gimenez et al. 2014). To take advantage of these developments, most statistical and modelling approaches are designed and first implemented using command line interfaces (e.g. Python and R), sometimes later being made available with a graphical user interface (GUI). Command line interfaces provide great flexibility and a precise record of what has been done, but are often difficult to interpret and tailor to specific data sets and analyses (Mislan et al. 2016). Conversely, GUIs are easier to navigate and extend the accessibility of the approach to many more users, but concessions are often made in terms of the availability of software that includes GUIs, customisability and reproducibility.

Species distribution modelling is a growing field in ecology. These models are widely used and often form the basis for important policy and conservation decisions. Due to their popularity, many SDM techniques have been made available with GUIs (sometimes referred to as 'turn-key' approaches) providing access to quite complex models by a much greater audience. These turn-key approaches are often referred to as 'black box' software, so named because many (or most) of the myriad adjustable SDM settings are concealed or inaccessible to the user. This is usually a deliberate decision, designed to increase the accessibility of the approach, but it can lead to users voluntarily or involuntarily ignoring important model parameterisations (Ahmed et al. 2015). What remains in question is the importance of the trade off between accessibility, customisability and reproducibility.

4.1 The realities of ecological modelling

Ecological models are mathematical objects intended to represent real-world phenomena. However, the world is fundamentally complex and it is probably impossible to capture this in any model (Evans 2012). In an attempt to make their models general, realistic and precise, modellers need to make decisions about which elements to include in their models and which to disregard. This almost always requires not only specialist modelling skills but also a detailed knowledge of the study species and its responses to the environment (Guisan et al. 2013). Also, ecological models are imperfect: an ecological model can only be as good as its input data. Uncertainty in model fit and its potential impact on predictions must be critically assessed and disclosed, when presented. The scarcity of these specialist skills has driven an increase in the availability of user-friendly web interfaces, such as the turn-key approaches evaluated here, but there is concern over the ability of inexperienced users and non-experts to adequately explore data sets and adjust model settings (where available) (Guisan et al. 2013).

4.2 Assessment of turn-key systems

In this report we evaluated four turn-key systems designed to provide access to SDM approaches by non-specialists: Lifemapper, the Atlas of Living Australia (ALA), the Biodiversity and Climate Change Virtual Laboratory (BCCVL), and *Wallace*. Aside from Lifemapper, which was not operational at the time of evaluation, most were straightforward to use and provided access to popular SDM approaches, and exporting results and spatial predictions was straightforward. *Wallace* was slightly more complicated than the ALA and BCCVL because it requires an active R session and two lines of code to initiate, but is still likely to be accessible to most researchers.

Some systems provided access to more SDM algorithms than others. The BCCVL provided access to 17 different algorithms, whereas the ALA only provided access to one algorithm (MaxEnt). The recently released *Wallace* only provided access to two algorithms (BIOCLIM and MaxEnt), but the developers have provided the architecture for the community to add more modules in the future (Kass et al. 2018a).

All three systems that were successfully evaluated provided access to the most popular SDM approach, MaxEnt, which also has its own GUI. However, its 'black-box' nature leaves many applications of it open to criticism (Merow et al. 2013; Ahmed et al. 2015). The level of customisability varied among turn-key systems. For example, many important settings such as the selection of feature shapes (which control the complexity of fitted species–environment relationships) could not be defined in the ALA (despite a strong recommendation that these be selected prior to model building: Elith et al. 2011; Merow et al. 2013), but could be defined in the BCCVL and *Wallace*.

All systems evaluated here had some capacity to download from public occurrence data networks, and species occurrence and spatial data could be directly uploaded by the user in all cases except for the ALA, where there is currently no option to upload spatial data. The ability for the user to upload their own data is important because it not only allows use of the tool across a range of countries and continents, but also allows the user to provide data most pertinent to their target species.

All turn-key systems evaluated included an ability to assess model fit, including fitted functions of environmental variables, and area under the receiver-operator curve (AUC) statistics, and sometimes they included a Jack-knife for assessing each environmental covariate's importance to the model (the ALA only). However, all outputs required model assessment skills and knowledge of the target species' ecology. The assessment of outliers from the primary data and the ability to assess the impact of data bias were limited across all systems.

The issue of reproducibility was also addressed in two systems: the BCCVL and *Wallace*. Both facilitated export of R code used to generate the results, enabling the user to re-run analyses and include the supplementary information of a research output. There is no option to export R-script in the ALA.

While it was unfortunate that Lifemapper was not operational at the time of this evaluation, it also highlighted another potential flaw with online turn-key modelling approaches such as the ALA and BCCVL: they are dependent on the web-hosting services of organisations that

may not reliably maintain them in the long term. This is likely to be especially relevant for organisations that rely on funding cycles to fund required web domains and servers, and to employ IT professionals with the responsibility for maintaining and updating the resources in perpetuity. Since *Wallace* is available as an open-source R-package and associated Shiny-based GUI, it may prove more reliable, especially if it becomes a popular resource that is expanded upon by the scientific community.

4.3 Development of a turn-key SDM system at MWLR

In 2015 a survey of MWLR staff on the use of SDM was carried out. The aim was to elicit requirements for a potential SDM platform to be established within MWLR. Data from 19 responses showed that roughly 30% of those surveyed had carried out some form of SDM work. The results also showed that these staff generally felt competent in the use of SDMs and understood the advantages and limitations of the approach. Issues frequently encountered were most often related to model selection, access to both species and environmental data, model evaluation, and computer performance. Most respondents (69%) ran their SDMs in R, indicating that a high proportion are comfortable with a command-line approach to ecological modelling.

From this survey it was concluded that the development of turn-key platforms was not required to meet the needs of MWLR staff. Rather, a number of needs were identified relating to: data/metadata access, formatting, standardisation, validation, and reproducible data-pipelining. Another need identified was the development of flexible solutions that can be adopted and tailored in familiar data-processing environments, such as R, to facilitate robust, documented and reproducible analyses.

The findings of this report support those of the 2015 survey in showing there are inherent problems with the turn-key SDM platform approach. While there have been some advances since their 2013 review, Guisan et al. (2013) found that key components of the model building process – such as evaluation of model fit and performance, uncertainty assessment and inspection of response curves – were not available in many turn-key SDM applications. They stated that they cannot advocate the use of potentially over-simplified tools, such as 'black box', turn-key modelling software, to support conservation decisions, and called for a wider recognition that SDMs should be developed by experts with intimate knowledge of the target species and statistical approaches.

Given the results of the 2015 MWLR survey, it is unlikely that the development of a turn-key approach will be of significant benefit to MWLR staff. While computer speed and data access issues could be addressed through running models on a centralised server, flexibility in model selection and evaluation – important issues identified in the survey – are constraints across all the turn-key systems evaluated.

4.4 Alternatives to turn-key SDM systems

There has been a strong push among ecologists to increase the use and understanding of computer code in ecology (Joppa et al. 2013; Mislan et al. 2016). Due to its popularity among ecologists, many of the cutting-edge developments in SDM are released in the code-based

computer program R (and, to a lesser extent, Python). While not as accessible as turn-key approaches, the use of these programs allows ecologists to run their analyses using the most up-to-date methods available, often long before they are made available with turn-key interfaces. Also, users are not restricted to the (often small) selection of SDM approaches made available by turn-key system developers, but instead can select from the growing range of algorithms being continually developed by the scientific community.

Within R there are several packages that can be used to run SDMs. One of the more commonly used packages is dismo, which includes functions for many popular SDM methods, has direct access to GBIF point records and BIOCLIM climate covariates, and has the ability to produce ensemble predictions (Hijmans et al. 2017). A recently developed package called sdm also provides an extensive framework for developing SDMs using several approaches, and also allows interrogation of results through a GUI (Naimi & Araújo 2016, 2018). There are specific packages for running SDM algorithms, such as maxnet for MaxEnt (Phillips Steven 2017; Phillips Steven et al. 2017), gbm for Boosted Regression Trees, and randomForest for Random Forests (Liaw & Wiener 2002; Breiman et al. 2018). These packages are under constant development by the user community, but all require greater experience with computer code and programs such as R than an equivalent turn-key system.

5 Conclusions and recommendations

Turn-key systems are relatively easy to use, produce appealing graphical outputs and extend the accessibility of SDMs to many more users. However, they result in serious concessions because many (or most) of the myriad adjustable SDM settings are concealed or inaccessible to the user.

We do not advocate that MLWR support the development and subsequent use of potentially over-simplified tools, such as 'black box', turn-key modelling software, by end-users to support conservation decisions. MLWR needs to promote a wider recognition that SDMs should be developed by experts with clear knowledge of the target species and statistical assumptions. This would be best achieved by continuing to support the development of in-house expertise in SDM and ensuring that outputs of SDM that can support decision-making are widely publicised and made available.

There is little interest among MWLR staff in using turn-key systems, so we do not recommend tailoring existing turn-key systems for use in New Zealand for internal MWLR purposes. A current barrier to developing credible SDMs in New Zealand is ready access to both species and environmental data.

MWLR is well placed to:

- further develop effective data delivery pipelines
- provide primary data for users to incorporate into their own modelling (e.g. make data available to be harvested by R packages in the same way the R-package dismo has direct access to GBIF point records)

- provide spatial covariate data for users to incorporate into their own modelling without a lot of pre-processing (e.g. allow automatic scaling of layers to consistent resolutions).

6 Acknowledgements

This work was funded under MBIE contract to Landcare Research PROP-38356-ETR-LCR. We thank Tom Etherington for reviewing the report and Ray Prebble for editing.

7 References

Ahmed SE, McInerny G, O'Hara K, Harper R, Salido L, Emmott S, Joppa LN 2015. Scientists and software – surveying the species distribution modelling community. Diversity and Distributions 21(3): 258–267.

Anderson BJ, Akçakaya HR, Araújo MB, Fordham DA, Martinez-Meyer E, Thuiller W, et al. 2009. Dynamics of range margins for metapopulations under climate change. Proceedings of the Royal Society B: Biological Sciences 276(1661): 1415–1420.

Atlas of Living Australia 2018. Atlas of Living Australia website. Retrieved 19 March 2018 from http://www.ala.org.au

Beaugrand G, Lenoir S, Ibañez F, Manté C 2011. A new model to assess the probability of occurrence of a species, based on presence-only data. Marine Ecology Progress Series 424: 175–190.

Besag J 1974. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society. Series B (Methodological) 36(2): 192–236.

Bolker BM 2008. Ecological models and data in R. Princeton, NJ, Princeton University Press.

Booth TH, Nix HA, Busby JR, Hutchinson MF 2014. bioclim: the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. Diversity and Distributions 20(1): 1–9.

Borregaard MK, Hart EM 2016. Towards a more reproducible ecology. Ecography 39(4): 349–353.

Breiman L, Cutler A, Liaw A, Wiener M 2018. randomForest: Breiman and Cutler's Random Forests for classification and regression, R package version 4.6-14.

Brown DG 1994. Predicting vegetation types at treeline using topography and biophysical disturbance variables. Journal of Vegetation Science 5(5): 641–656.

Busby JR 1991. BIOCLIM: a bioclimatic analysis and prediction system. In: Margules CR, Austin MP eds. Nature conservation: cost effective biological surveys and data analysis. Australia, CSIRO. Pp. 64–68.

Carpenter G, Gillison AN, Winter J 1993. DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. Biodiversity & Conservation 2(6): 667–680.

Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J 2017. shiny: web application framework for R, R package version 1.0.3.

Civco DL 1993. Artificial neural networks for land-cover classification and mapping. International Journal of Geographical Information Systems 7(2): 173–186.

Drake JM, Randin C, Guisan A 2006. Modelling ecological niches with support vector machines. Journal of Applied Ecology 43(3): 424–432.

El-Gabbas A, Dormann CF 2018. Improved species-occurrence predictions in data-poor regions: using large-scale data and bias correction with down-weighted Poisson regression and Maxent. Ecography 41(7): 1161–1172.

Elith J, Leathwick JR 2009. Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics 40: 677–697.

Elith J, Franklin J 2013. Species distribution modeling. In: Levin SA ed. Encyclopedia of biodiversity. 2nd edition. Amsterdam, Academic Press. Pp. 692–705.

Elith J, Graham CH, Anderson RP, Dudík M, Ferrier S, Guisan A, et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29(2): 129–151.

Elith J, Graham CH 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. Ecography 32(1): 66–77.

Elith J, Leathwick JR, Hastie T 2008. A working guide to boosted regression trees. Journal of Animal Ecology 77(4): 802–813.

Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17(1): 43–57.

Evans MR 2012. Modelling ecological systems in a changing world. Philosophical Transactions of the Royal Society B: Biological Sciences 367(1586): 181–190.

Fensham RJ, Fairfax RJ, Butler DW, Bowman DMJS 2003. Effects of fire and drought in a tropical eucalypt savanna colonized by rain forest. Journal of Biogeography 30(9): 1405–1414.

Ferrier S, Guisan A 2006. Spatial modelling of biodiversity at the community level. Journal of Applied Ecology 43(3): 393–404.

Franklin J 2002. Enhancing a regional vegetation map with predictive models of dominant plant species in chaparral. Applied Vegetation Science 5(1): 135–146.

Franklin J 2009. Mapping species distributions: spatial inference and prediction. Cambridge, UK, Cambridge University Press.

German GWH, Gahegan MN 1996. Neural network architectures for the classification of temporal image sequences. Computers & Geosciences 22(9): 969–979.

Gimenez O, Buckland ST, Morgan BJT, Bez N, Bertrand S, Choquet R, et al. 2014. Statistical ecology comes of age. Biology Letters 10(12): 20140698.

Gu W, Swihart RK 2004. Absent or undetected?: effects of non-detection of species occurrence on wildlife–habitat models. Biological Conservation 116(2): 195–203.

Guillera-Arroita G, Lahoz-Monfort JJ, Elith J, Gordon A, Kujala H, Lentini PE, et al. 2015. Is my species distribution model fit for purpose? Matching data and models to applications. Global Ecology and Biogeography 24(3): 276–292.

Guisan A, Edwards TC, Hastie T 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling 157(2): 89–100.

Guisan A, Tingley R, Baumgartner JB, Naujokaitis-Lewis I, Sutcliffe PR, Tulloch AIT, et al. 2013. Predicting species distributions for conservation decisions. Ecology Letters 16(12): 1424–1435.

Guisan A, Zimmermann NE 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135(2): 147–186.

Guo Q, Kelly M, Graham CH 2005. Support vector machines for predicting distribution of Sudden Oak Death in California. Ecological Modelling 182(1): 75–90.

Hallgren W, Beaumont L, Bowness A, Chambers L, Graham E, Holewa H, et al. 2016. The Biodiversity and Climate Change Virtual Laboratory: where ecology meets big data. Environmental Modelling & Software 76: 182–186.

Hampton SE, Anderson SS, Bagby SC, Gries C, Han X, Hart EM, et al. 2015. The Tao of open science for ecology. Ecosphere 6(7): 120.

Hastie T, Tibshirani R 1987. Generalized additive models: some applications. Journal of the American Statistical Association 82(398): 371–386.

Hastie T, Tibshirani R 1996. Discriminant analysis by Gaussian mixtures. Journal of the Royal Statistical Society. Series B (Methodological) 58(1): 155–176.

Hastie T, Tibshirani R, Buja A 1994. Flexible discriminant analysis by optimal scoring. Journal of the American Statistical Association 89(428): 1255–1270.

Hastie T, Tibshirani R, Friedman J 2001. The elements of statistical learning: data mining, inference and prediction. New York, NY, Springer.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25(15): 1965–1978.

Hijmans RJ, Phillips SJ, Leathwick JR, Elith J 2017. dismo: species distribution modelling, R package version 1.1-4.

Hirzel AH, Hausser J, Chessel D, Perrin N 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology 83(7): 2027–2036.

Jager MM, Richardson SJ, Bellingham PJ, Clearwater MJ, Laughlin DC 2015. Soil fertility induces coordinated responses of multiple independent functional traits. Journal of Ecology 103(2): 374–385.

Joppa LN, McInerny G, Harper R, Salido L, Takeda K, O'Hara K, et al. 2013. Troubling trends in scientific software use. Science 340(6134): 814–815.

Kass JM, Vilela B, Aiello-Lammens ME, Muscarella R, Merow C, Anderson RP 2018a. Wallace: a flexible platform for reproducible modeling of species niches and distributions built for community expansion. Methods in Ecology and Evolution 9(4): 1151–1156.

Kass JM, Vilela B, Aiello-Lammens ME, Muscarella R, Merow C, Anderson RP 2018b. Wallace: a modular platform for reproducible modeling of species niches and distributions, R package version 1.0.4.

Kearney M, Porter W 2009. Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. Ecology Letters 12(4): 334–350.

Leathwick J, Morgan F, Wilson G, Rutledge D, McLeod M, Johnson K 2002. Land environments of New Zealand: a technical guide. Wellington, Ministry for the Environment.

Leathwick JR, Elith J, Hastie T 2006. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. Ecological Modelling 199(2): 188–196.

Liaw A, Wiener M 2002. Classification and regression by randomForest. R News 2(3): 18–22.

Linderman M, Liu J, Qi J, An L, Ouyang Z, Yang J, et al. 2004. Using artificial neural networks to map the spatial distribution of understorey bamboo from remote sensing data. International Journal of Remote Sensing 25(9): 1685–1700.

MacKenzie DI, Nichols JD, Lachman GB, Droege S, Andrew RJ, Langtimm CA 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83(8): 2248–2255.

McCullagh P, Nelder JA 1983. Generalized linear models. London, UK, Chapman and Hall.

McGlone MS 1989. The Polynesian settlement of New Zealand in relation to environmental and biotic changes. New Zealand Journal of Ecology 12: 115–129.

McGlone MS, Richardson SJ, Jordan GJ 2010. Comparative biogeography of New Zealand trees: species richness, height, leaf traits and range sizes. New Zealand Journal of Ecology 34(1): 137–151.

Merow C, Silander JA 2014. A comparison of Maxlike and Maxent for modelling species distributions. Methods in Ecology and Evolution 5(3): 215–225.

Merow C, Smith MJ, Silander JA 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography 36(10): 1058–1069.

Miller DA, Nichols JD, McClintock BT, Grant EHC, Bailey LL, Weir LA 2011. Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. Ecology 92(7): 1422–1428.

Mislan KAS, Heer JM, White EP 2016. Elevating the status of code in ecology. Trends in Ecology & Evolution 31(1): 4–7.

Monserud RA, Leemans R 1992. Comparing global vegetation maps with the Kappa statistic. Ecological Modelling 62(4): 275–293.

Muñoz J, Felicísimo ÁM 2004. Comparison of statistical methods commonly used in predictive modelling. Journal of Vegetation Science 15(2): 285–292.

Naimi B, Araújo MB 2016. sdm: a reproducible and extensible R platform for species distribution modelling. Ecography 39(4): 368–375.

Naimi B, Araújo MB 2018. sdm: species distribution modelling, R package version 1.0-46.

Nix HA 1986. A biogeographic analysis of Australian elapid snakes. In: Longmore R ed. Atlas of elapid snakes of Australia. Canberra, Australian Government Publishing Service. Pp. 4–15.

Olden JD, Lawler JJ, Poff NL 2008. Machine learning methods without tears: a primer for ecologists. Quarterly Review of Biology 83(2): 171–193.

Ovaskainen O, Roy DB, Fox R, Anderson BJ 2016. Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. Methods in Ecology and Evolution 7(4): 428–436.

Phillips SJ 2017. maxnet: fitting 'Maxent' species distribution models with 'glmnet', R package version 0.1.2.

Phillips SJ, Anderson RP, Dudík M, Schapire RE, Blair ME 2017. Opening the black box: an open-source release of Maxent. Ecography 40(7): 887–893.

Phillips SJ, Anderson RP, Schapire RE 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190(3): 231–259.

Phillips SJ, Dudík M 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31(2): 161–175.

Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, et al. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19(1): 181–197.

Prasad AM, Iverson LR, Liaw A 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9(2): 181–199.

R Core Team 2016. R: a language and environment for statistical computing. Vienna, R Foundation for Statistical Computing.

Renner IW, Warton DI 2013. Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. Biometrics 69(1): 274–281.

Richardson SJ, Peltzer DA, Allen RB, McGlone MS, Parfitt RL 2004. Rapid development of phosphorus limitation in temperate rainforest along the Franz Josef soil chronosequence. Oecologia 139(2): 267–276.

Rocchini D, Hortal J, Lengyel S, Lobo JM, Jiménez-Valverde A, Ricotta C, et al. 2011. Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. Progress in Physical Geography: Earth and Environment 35(2): 211–226.

Royle JA, Chandler RB, Yackulic C, Nichols JD 2012. Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. Methods in Ecology and Evolution 3(3): 545–554.

Stockwell D, Peters D 1999. The GARP modelling system: problems and solutions to automated spatial prediction. International Journal of Geographical Information Science 13(2): 143–158.

Stockwell DRB, Beach JH, Stewart A, Vorontsov G, Vieglais D, Pereira RS 2006. The use of the GARP genetic algorithm and Internet grid computing in the Lifemapper world atlas of species biodiversity. Ecological Modelling 195(1): 139–145.

Stockwell DRB, Noble IR 1992. Induction of sets of rules from animal distribution data: a robust and informative method of data analysis. Mathematics and Computers in Simulation 33(5): 385–390.

Stokland JN, Halvorsen R, Støa B 2011. Species distribution modelling–Effect of design and sample size of pseudo-absence observations. Ecological Modelling 222(11): 1800–1809.

Stoklosa J, Daly C, Foster SD, Ashcroft MB, Warton DI 2015. A climate of uncertainty: accounting for error in climate variables for species distribution models. Methods in Ecology and Evolution 6(4): 412–423.

VanDerWal J, Shoo LP, Graham C, Williams SE 2009. Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know? Ecological Modelling 220(4): 589–594.

Ver Hoef JM, Peterson EE, Hooten MB, Hanks EM, Fortin MJ 2018. Spatial autoregressive models for statistical inference from ecological data. Ecological Monographs 88(1): 36–59.

Wardle P 1991. Vegetation of New Zealand. Cambridge, UK, Cambridge University Press.

Warton DI, Blanchet FG, O'Hara RB, Ovaskainen O, Taskinen S, Walker SC, et al. 2015. So many variables: Joint modeling in community ecology. Trends in Ecology & Evolution 30(12): 766–779.

Warton DI, Shepherd LC 2010. Poisson point process models solve the 'pseudo-absence problem' for presence-only data in ecology. Annals of Applied Statistics 4(3): 1383–1402.

Wells A, Duncan RP, Stewart GH 2001. Forest dynamics in Westland, New Zealand: the importance of large, infrequent earthquake-induced disturbance. Journal of Ecology 89(6): 1006–1018.

Wintle BA, Elith J, Potts JM 2005. Fauna habitat modelling and mapping: a review and case study in the Lower Hunter Central Coast region of NSW. Austral Ecology 30(7): 719–738.

Wright SJ 1992. Seasonal drought, soil fertility and the species density of tropical forest plant communities. Trends in Ecology & Evolution 7(8): 260–263.

Yee TW, Mitchell ND 1991. Generalized additive models in plant ecology. Journal of Vegetation Science 2(5): 587–602.

Zimmermann NE, Kienast F 1999. Predictive mapping of alpine grasslands in Switzerland: species versus community approach. Journal of Vegetation Science 10(4): 469–482.