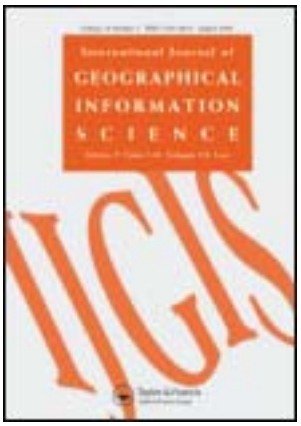

International Journal of Geographical Information Science

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tgis20

Which environmental variables should I use in my biodiversity model?

Kristen J. Williams [a] , Lee Belbin [b] , Michael P. Austin [a] , Janet L. Stein [c] & Simon Ferrier [a]

[a] Ecosystem Sciences, CSIRO, Canberra, Australia

[b] Atlas of Living Australia, Hobart, Australia

[c] Fenner School of Environment and Society, Australian National University, Canberra, Australia

Version of record first published: 09 Jul 2012.

To cite this article: Kristen J. Williams , Lee Belbin , Michael P. Austin , Janet L. Stein & Simon Ferrier (2012): Which environmental variables should I use in my biodiversity model?, International Journal of Geographical Information Science, 26:11, 2009-2047

To link to this article: http://dx.doi.org/10.1080/13658816.2012.698015

International Journal of Geographical Information Science
Vol. 26, No. 11, November 2012, 2009–2047

Which environmental variables should I use in my biodiversity model?

Kristen J. Williams[a]*, Lee Belbin[b], Michael P. Austin[a], Janet L. Stein[c] and Simon Ferrier[a]

[a]*Ecosystem Sciences, CSIRO, Canberra, Australia;* [b]*Atlas of Living Australia, Hobart, Australia;* [c]*Fenner School of Environment and Society, Australian National University, Canberra, Australia*

(*Received 14 January 2012; final version received 21 May 2012*)

Appropriate selection of environmental variables is critical to the performance of biodiversity models, but has received less attention than the choice of modelling method. Online aggregators of biological and environmental data, such as the Global Biodiversity Information Facility and the Atlas of Living Australia, necessitate a rational approach to variable selection. We outline a set of general principles for systematically identifying, compiling, evaluating and selecting environmental variables for a biodiversity model. Our approach aims to maximise the information obtained from the analysis of biological records linked to a potentially large suite of spatial environmental variables. We demonstrate the utility of this structured framework through case studies with Australian vascular plants: regional modelling of a species distribution, continent-wide modelling of species compositional turnover and environmental classification. The approach is informed by three components of a biodiversity model: (1) an ecological framework or conceptual model, (2) a data model concerning availability, resolution and variable selection and (3) a method for analysing data. We expand the data model in structuring the problem of choosing environmental variables. The case studies demonstrate a structured approach for the: (1) cost-effective compilation of variables in the context of an explicit ecological framework for the study, attribute accuracy and resolution; (2) evaluation of non-linear relationships between variables using knowledge of their derivation, scatter plots and dissimilarity matrices; (3) selection and grouping of variables based on hypotheses of relative ecological importance and perceived predictor effectiveness; (4) systematic testing of variables as predictors through the process of model building and refinement and (5) model critique, inference and synthesis using direct gradient analysis to evaluate the shape of response curves in the context of ecological theory by presenting predictions in both geographic and environmental space.

Keywords: environmental variables; Atlas of Living Australia; species distribution modelling; MaxEnt; generalised dissimilarity modelling; association analysis; classification; guidelines

1. Introduction

There is an increasing trend towards free and open access to primary biodiversity and environmental data via web servers for use in ecological research, natural resources management and conservation decision-making. Data aggregators such as the Global Biodiversity Information Facility (GBIF; http://www.gbif.org/) and the Atlas of Living

Target Journal: Second Special Issue on Spatial Ecology, International Journal of Geographical Information Sciences, edited by Shawn Laffan, Andy Skidmore and Janet Franklin.
*Corresponding author. Email: kristen.williams@csiro.au

ISSN 1365-8816 print/ISSN 1362-3087 online

http://dx.doi.org/10.1080/13658816.2012.698015
http://www.tandfonline.com

Australia (ALA; http://www.ala.org.au/) facilitate the use of information about the occurrence of organisms. In addition, various national and global networks collect, manage and disseminate a wide range of environmental data (Hijmans *et al*. 2005, Hutchinson *et al*. 2008, Gallant and Read 2009, Minasny and McBratney 2010). This increase in data availability has been accompanied by open-access software to characterise the patterns and trends in biological and environmental data (e.g. generalised dissimilarity modelling (GDM, Ferrier *et al*. 2007), MaxEnt (Phillips and Dudík 2008) and Biodiverse (Laffan *et al*. 2010)). The spatial portal of the Atlas of Living Australia (http://spatial.ala.org.au/), for example, enables analysts to use millions of taxon observations to predict distributions or build an environmental domain classification from over 250 environmental layers (see http://dashboard.ala.org.au/dashboard/). These variables describe globally relevant facets of climate, soil and terrain. However, ready access to data and analytical tools does not solve the problem of deciding which environmental variables are appropriate for use in a biodiversity model. Appropriate selection of environmental variables is critical to the performance of the model and its potential application in prediction and explanation (Austin 2002b, Elith and Leathwick 2009), but has received less attention than the choice of modelling method (Franklin 2009).

Making decisions about which environmental variables to use and their relative contribution remains a challenge for species distribution modelling (Araújo and Guisan 2006, Peterson and Nakazawa 2008, Franklin 2009, Ashcroft *et al*. 2011, Austin and Van Niel 2011, Peterson *et al*. 2011). Some researchers may purposefully limit their analyses *a priori* to a few justified or easily measured predictors (Austin *et al*. 1990) with the resultant potential for under-specified models. Others may draw upon a more comprehensive set of variables (Williams *et al*. 2000) but risk over-fitting. Because the selection of predictor variables is a critical step in modelling species distributions, Araújo and Guisan (2006) suggested that more attention should be given to the explanatory power and ecological basis for choosing variables. They and other workers emphasised the importance of selecting variables that are physiologically relevant (Franklin 1995, Austin 2007, Austin and Van Niel 2011).

Selecting environmental variables is one component of a structured approach to biodiversity distribution modelling that comprises (1) an ecological framework or conceptual model, which includes the theory used to link environmental predictors to biodiversity distribution; (2) a data model, which considers data availability, resolution and selection, and (3) a statistical model, which includes the modelling method and the selection of explanatory variables during model building and evaluation (Austin 2002b).

In this article, we first review how these three components interact to influence the selection of spatial environmental variables through an expansion of the framework for a data model. We then demonstrate this structured approach to identifying, compiling, evaluating and selecting environmental variables through three typical applications in vegetation science: (1) regional modelling of a species distribution, (2) continent-wide modelling of spatial turnover in species composition and (3) environmental classification. We conclude by outlining a set of general principles to guide future applications.

2. A structured approach to biodiversity distribution modelling

2.1. The ecological model or conceptual framework

Conceptual underpinnings in ecology are prerequisites for deciding what facets of environment are likely to be of interest, given a particular taxon and modelling purpose (Austin 2007). For example, seven groups of variables are usually associated with

vegetation responses: light, temperature, nutrients, water, CO_2, disturbance and biota (Austin and Van Niel 2011). The complexity of relationships between environmental gradients and plant distribution are described diagrammatically by Guisan and co-workers (Guisan and Zimmermann 2000, Guisan and Thuiller 2005). Drawing on the continuum concept in vegetation ecology, this framework distinguishes factors that are functionally relevant to the physiology of a species and its fundamental niche and other ecological constraints such as species interactions, disturbance regimes and biogeographic barriers that limit species occurrence. A wide range of environmental conditions and biotic interactions act together to determine the dynamic trade-offs in growth and development processes, at the physiological and molecular levels, thereby influencing the occurrence and behaviour of species at the individual and population level (Smith and Huston 1989). These conceptual frameworks provide a checklist of the broad groups of environmental factors to be considered '*a priori*' (i.e. knowledge is justified by arguments of a certain kind) for inclusion in a model and distinguish variables that are either proximal or distally related to the resources and conditions controlling growth, reproduction, morphology and behaviour (Austin 2005, Franklin 2009).

2.2. The data model

The data model considers the response and explanatory data that are used for a given purpose. Biodiversity distribution models typically use the presence–absence data from systematic surveys, the presence-only data from aggregations of opportunistic observations and/or the records from museum or herbarium collections. For example, generalised additive modelling of species distributions is an appropriate technique for presence–absence data, whereas MaxEnt is better used with presence-only data (Elith *et al*. 2011). In species composition (community-level) analyses, such as applications of GDM (Ferrier *et al*. 2007), the presence–absence data are preferred (Ferrier and Guisan 2006). However, the relatively large amount of presence-only data from aggregators, such as the GBIF, can reduce the effects of geographic and environmental bias, allowing its wider use as a proxy for presence–absence data in analysing the environmental determinants of species composition (Kent and Carmel 2011). Irrespective of the biodiversity response variables (species occurrence, abundance and compositional dissimilarity), the same conceptual framework applies for selecting environmental variables relevant to a study. We recognize four stages in the process of selecting environmental variables: identifying, compiling, evaluating and testing the fit (model building), as outlined in Figure 1.

2.2.1. Identifying the resolution and extent of spatial environmental variables

The outcomes of an analysis are critically affected by the resolution of the data and the geographic extent of the study area. Resolution and extent are considerations in searching for environmental variables relevant to the study's purpose (Figure 1). Despite well-established guidelines for selecting an ecological analysis region (Austin *et al*. 1996, Anderson and Raza 2010), many studies still define extent by jurisdiction or planning boundaries. To avoid truncated model predictions, the analysis region should include the known environmental or geographic limits of the study taxon. For example, Williams and Potts (1996) mapped the broad geographic range of *Eucalyptus* species in Tasmania as a basis for systematically defining the ecological analysis region used in generalised linear models of individual species' distributions (applied in Williams (1998)). Similarly, MaxEnt requires the landscape of interest to be defined by the ecologist as encompassing the expected predicted range of the species (Elith *et al*. 2011).

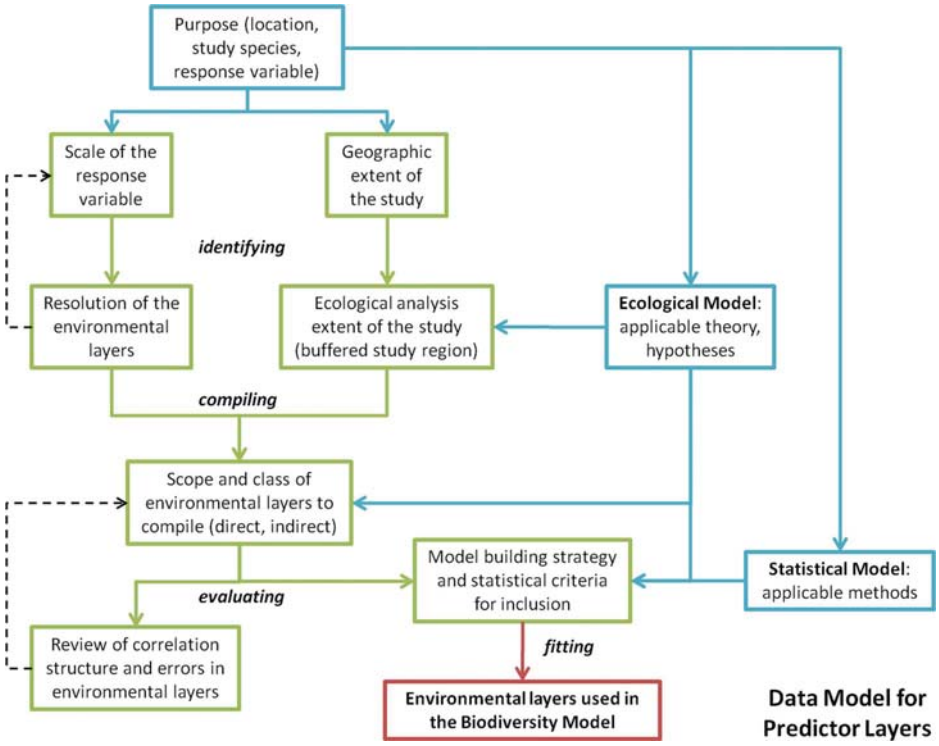

Figure 1. An expansion of the data model within the framework for spatial prediction illustrating the relationship between the data model, ecological model and statistical model for the purpose of selecting environmental variables for use in a biodiversity model. The data model is central to the process and interacts with the other framework model components (blue boxes). The overarching purpose of the study determines the underpinning ecological theory, data and statistical model. The green boxes represent stages in the process of selecting environmental layers grouped into four principle activities, in order: identifying, compiling, evaluating and selecting environmental variables. The red box highlights the outcome of the activities. Dotted lines indicate activities that may be conducted iteratively or where feedback affects a previous decision.

Although the purpose of the study and observation scale of the response variable dictate the desired spatial resolution of environmental variables, a lack of suitable resolution for explanatory data and sparse sampling of response data both influence the predictive resolution of the model (dotted lines in Figure 1). Ideally, the scale of the response and explanatory data should match closely (Graf *et al*. 2005, Guisan *et al*. 2007, Kaliontzopoulou *et al*. 2008). In practice, however, environmental variables of varying resolution are compiled and resampled to a common spatial grid for modelling (Elith and Leathwick 2009). The resolution of the grid aims to balance the resolution of key explanatory variables and the observation scale of the response.

The predictive resolution of the response then depends on the representativeness and intensity of sampling. Sparsely sampled regions generally contain less information about the relationship between response and explanatory variables than more densely sampled regions. In these situations, fewer variables may be required to adequately predict the regions of broadly similar environments, but will inadequately define the range limits (Pearson *et al*. 2007). While some analysts may consider coarsening the spatial resolution of environmental variables to match the information content of sparsely sampled regions

prior to model building, a more effective approach is to develop a model using the best available data and *post hoc* determine the classification resolution of the prediction map using spatial statistics (Hagen-Zanker 2009).

2.2.2. *Scope of environmental predictors to compile for a study*

Considering the large number of environmental variables potentially relevant to a study, the challenge is to find appropriate sets of proximal variables consistent with the knowledge of biophysical process and study resolution (Franklin 2009). The first step is to review what is known about the ecology and physiology of the biodiversity of interest to inform hypotheses describing the relationship between the response and a desired set of explanatory variables.

Figure 2, adapted from Guisan and Zimmermann (2000), shows the compilation of predictors for a vegetation modelling project. Arrows in the figure show how particular

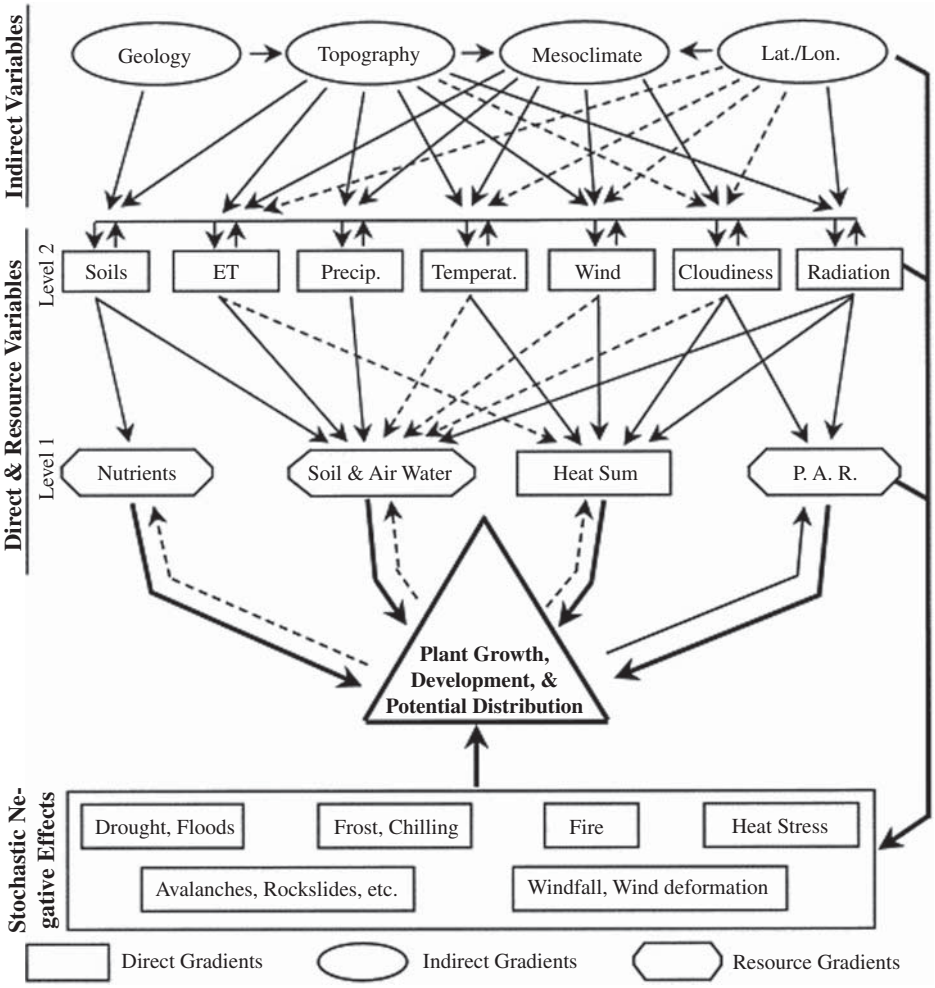

Figure 2. Example of a conceptual model of relationships between resources, direct and indirect environmental gradients and their influence on growth, performance and geographical distribution of vascular plants and vegetation. Reproduced with permission from Guisan and Zimmermann (2000; Figure 3).

indirect variables interact to generate more direct environmental drivers through biophysical process models: nutrients, soil and atmospheric water balance, thermal development conditions and photosynthetically active radiation. Plant distributions are also influenced by stochastic processes such as extreme heat or cold, landslip or erosion, high winds, drought, flood and fire. Therefore, environmental variables that directly influence a wide range of climatic and disturbance regimes via their rate, intensity, duration and frequency may be required to effectively model distribution patterns. Direct and indirect proxies for these variables can be compiled from existing sources or generated by modelling these processes.

Physiologically relevant environmental measures are expected to lead to more robust ecological models and may justify the effort required to model these processes. For example, Leathwick and Whitehead (2001) found that derivation of more direct measures of air and soil water deficit improved models of New Zealand tree species. Battaglia and Williams (1996) demonstrated that common measurement units for soil moisture distinguished the ecotonal gradient between two co-occurring *Eucalyptus* species at both physiological and landscape scales.

An environmental variable may act both as a resource that provides building blocks for growth processes and as a condition that fulfils the requirements for physiological processes to function effectively. For example, solar radiation, in addition to being a resource that determines primary productivity, plays an important role as a condition in regulating or controlling the phases of growth, development and defence in plants through various forms of light quality perception and wavelength signalling (Kazan and Manners 2011).

Various measures of geographic and topographic positions (latitude, longitude, altitude, slope and aspect) represent indirect environmental gradients that influence biodiversity via correlated proximal variables (Austin 2002b). Models that include such indirect variables along with direct variables may be justified where biophysical process models are incompletely specified, critical data are missing for particular factors or issues of attribute accuracy and spatial resolution limit their effectiveness. For example, Ashcroft *et al*. (2008) developed exposure indices based on topographic protection to serve as a proxy for the effect of seasonally prevailing winds on local soil moisture and humidity in the absence of the proximal predictor. Similarly, maps of surface geology (lithology) may be used even though indirectly related to aspects of the soil environment via the water balance, nutrient availability, plant support and root exploration. Assuming they correlate consistently with direct gradients, indirect predictors or their proximal derivatives may be justified in the absence of more direct estimates in a biodiversity model.

2.2.3. Evaluating the environmental predictors that are used in model building

The common physical and biological processes that relate candidate environmental predictors provide insight into their multivariate correlations (Figure 2). Potentially redundant variables can be distinguished from relatively independent variables using a correlation analysis that accounts for non-linearity in pair-wise relationships. For example, Elith *et al*. (2010; Appendix s4) used 18 agro-climatic regions as defined by Hutchinson *et al*. (2005) to account for spatially non-linear patterns in pair-wise Pearson's correlation coefficients. The different sets of correlation coefficients in each region informed the choice of relatively independent variables used in the respective species distribution models. This process is informative but not definitive as pair-wise correlations between environmental

variables do not necessarily reveal the most effective environmental predictors because response–explanatory relationships are typically non-linear (Austin and Smith 1989).

2.3. Testing the fit of variables using a statistical model

With a potentially large number of alternative combinations of variables to consider, a structured framework for testing significance and eliminating variables that perform poorly is essential. Araújo and Guisan (2006) identified improved model selection and predictor contribution as one of the significant challenges in species distribution modelling. The process of fitting a statistical model provides both a test of the utility of candidate environmental variables and feeds back into the design of the data model (dotted lines in Figure 1).

Statisticians have developed different approaches to the problem of selecting significant predictors in regression-based modelling and evaluating the overall model performance (Hosmer and Lemeshow 1989). The common approaches to variable selection include backward, forward and stepwise procedures, and some lesser known techniques such as stage-wise (Hastie *et al*. 2007), best subset (King 2003) and purposeful (Bursac *et al*. 2008). However, automated procedures do not necessarily select the best set of explanatory variables, but a best subset based on the algorithm and set of criteria used (Pearce and Ferrier 2000). Other statistical or data mining techniques such as boosted regression trees can also assist with the selection of relevant variables (Elith *et al*. 2008, Magness *et al*. 2008), as can repeated permutations testing the match between response and explanatory variables.

In our case studies, relatively independent subgroups of correlated variables are tested separately before combining them, in an *a priori* order of importance, in a model. Our framework guides the identification of direct/proximal variables that are initially tested followed by tests for the additional effect of indirect variables. Therefore, identification of redundant variables relies on a combination of *a priori* ecological considerations, knowledge of the derivation and accuracy of each variable, awareness of relationships among variables and a rigorous process of testing the utility of alternative sets of predictors in a statistical model.

3. Methods

3.1. Case study context

We selected vascular plants as our study taxa and applied three ecological analysis methods as case studies: a species distribution model for *Eucalyptus delegatensis* R.T. Baker (family: Myrtaceae) using MaxEnt software (Phillips *et al*. 2006, Phillips and Dudík 2008); a model of spatial turnover in vascular plant species composition using GDM software (Ferrier *et al*. 2007) and an environmental domain analysis based on a non-hierarchical classification using PATN software (Belbin 1987; http://www.patn.com.au). We chose MaxEnt and GDM as these are robust methods for modelling presence-only response data (Elith *et al*. 2006, 2011), sourced from aggregators such as the Atlas of Living Australia.

Our study area is the continent of Australia and proximate islands with a land area of 7.7 million square kilometres. Australia is mostly warm to hot and very dry with seasonally wet or dry climates. A winter-dominated rainfall regime occurs in the south and a summer-dominated regime occurs in the north. More than 80% of the continent has at least 3 months each year without effective precipitation (Hutchinson *et al*. 2005).

3.2. Environmental variables

We limited our investigation to the comprehensive set of environmental variables, compiled by Williams *et al*. (2010), available online via the Atlas of Living Australia (see http://spatial.ala.org.au/layers). The variables represent the best available, nationally consistent 0.01°(∼1 km) resolution-gridded sources of climate, soil, geology and terrain information as of November 2009 (summarised in Table 1).

Table 1. List of environmental variables, labels and units.

Group	Data set	Label
Water	RAINI, RAINX	Minimum (I) and maximum (X) of monthly rainfall (mm)
	RPRECMIN, RPRECMAX	Minimum (MIN) and maximum (MAX) of monthly rainfall difference between successive months (mm/day)
	EVAPI, EVAPX	Minimum (I) and maximum (X) of monthly evaporation, averaged over 25 years centred on 1982 (mm)
	ARID_MIN, ARID_MAX	Minimum and maximum of monthly aridity index as ratio of precipitation to evaporation (dimensionless)
	ADEFI, ADEFX	Minimum (I) and maximum (X) of monthly precipitation deficit as precipitation minus evaporation (mm)
	SRAIN1MP, SRAIN2MP	Solstice (1) or equinox (2) rainfall seasonality ratio (dimensionless)
	SLRAIN1, SLRAIN2	Solstice (1) or equinox (2) rainfall seasonality factor index
Energy	RADNI, RADNX	Minimum (I) and maximum (X) of monthly mean rainfall-modified (cloudiness) solar radiation, averaged over 25 years centred on 1982 ($MJ/m^2/day$)
	MINTI, MINTX	Minimum (I) and maximum (X) of monthly mean minimum temperature (°C)
	MAXTI, MAXTX	Minimum (I) and maximum (X) of monthly mean maximum temperature (°C)
	TMINSABSI, TMAXABSX	Absolute minimum (I) and maximum (X) of daily temperature per month, averaged over 50 years centred on 1975 (°C)
	RTIMIN, RTIMAX	Minimum (MIN) and maximum (MAX) of monthly mean difference in minimum temperatures between successive months (°C/day)
	RTXMIN, RTXMAX	Minimum (MIN) and maximum (MAX) of monthly mean difference in maximum temperatures between successive months (°C/day)
	TRNGI, TRNGX	Minimum (I) and maximum (X) of monthly mean diurnal temperature range (°C)
	RH2MIN, RH2MAX	Minimum (MIN) and maximum (MAX) of monthly mean relative humidity, averaged over 25 years centred on 1982 (%)
	VPD2MIN, VPD2MAX	Minimum (MIN) and maximum (MAX) of monthly mean vapour pressure deficit, averaged over 25 years centred on 1982 (kPa)
	WINDSPMIN, WINDSPMAX	Minimum (MIN) and maximum (MAX) of monthly mean wind speed at 9 am or 3 pm, averaged over 25 years centred on 1982 (m/s)

(*Continued*)

Table 1. (*Continued*).

Group	Data set	Label
	WINDRMIN, WINDRMAX	Minimum (MIN) and maximum (MAX) of monthly wind run, averaged over 25 years centred on 1982 (km/day)
Soil	DATASUPT	Data levels supporting soil property interpretations (index)
	SOLDEPTH	Solum depth (surface and subsoil layers) (metres)
	SOLPAWHC	Plant-available soil water-holding capacity estimated from soil depth and clay content (mm)
	WR_UNR	Solum average unreliable water retention parameters (index)
	KSAT	Solum average median horizon saturated hydraulic conductivity (mm/h)
	KS_ERR	Solum average uncertainty of horizon saturated hydraulic conductivity estimates (index)
	CALCRETE	Calcrete in or below soil profile (presence)
	HPEDALITY	Hydrological scoring of pedality (score)
	COARSE	Soils dominated by coarse fragments including ironstone (class)
	CLAY	Solum average median clay content (%)
	BDENSITY	Solum average bulk density (mg/m^3)
	NUTRIENTS	Gross nutrient status (rating)
Geoscience	FERT	Inherent rock fertility (rating)
	GEOLLMEANAGE; GEOLLRNGEAGE	Geological age (log_{10}) mean and range (log_{10} M years)
	GRAVITY	Bouguer gravity anomalies (acceleration, Gal)
	MAGNETICS	Magnetic anomalies (nanoTesla, nT)
	WII_WGS1KB	Weathering intensity index
Terrain	SLOPE	Terrain slope (%)
	RELIEF	Terrain relief (metres)
	ROUGHNESS	Terrain roughness (%)
	TWI	Topographic wetness index (index)
	MRVBF	Valley bottom flatness (index)
	MRRTF	Ridgetop flatness (index)
	VALLEYBOTTOM	Local neighbourhood proportion valley bottoms (%)
	RIDGETOPFLAT	Local neighbourhood proportion ridge tops (%)
	EROSIONAL	Local neighbourhood proportion erosional surfaces (%)

Notes: Metadata for each variable can be viewed from the links provided at http://spatial.ala.org.au/layers. The full association matrix can be downloaded from http://spatial.ala.org.au/files/inter_layer_association.csv. Climate variables are long-term averages approximately centred on 1960, unless otherwise stated.

Climate is represented by 15 measures – rainfall, pan evaporation, precipitation deficit, aridity, minimum and maximum temperatures, diurnal temperature range, solar radiation (adjusted by rainfall as a surrogate for cloud cover), relative humidity, vapour pressure deficit, wind speed, wind run and rates of rise and fall in rainfall and in minimum and maximum temperatures. Each of these measures is represented by 12 long-term, mean monthly values as derived from ANUCLIM, software version 5.1 (Hutchinson *et al*. 2000). We evaluated the minimum and maximum monthly values of the annual variation because these indices reflect distinct seasonal means, are directly related to plant response and are less correlated with each other than either is with the annual mean. We included two compound measures of water as potential substitutes for rainfall and evaporation: precipitation deficit

(Harmsen *et al*. 2009) and aridity index (UNEP 1992 cited in Middleton and Thomas 1997). Two alternative measures of rainfall seasonality, both based on the ratio of summer to winter or spring to autumn rainfall, were also included. One is a factor using the logarithm of rainfall (detailed in Williams *et al*. (2010)) and the other a simple ratio (based on Austin (1998)). We also include a measure of mean absolute monthly temperature derived from 5 km gridded daily climate estimates for Australia (Jeffrey *et al*. 2001).

Substrate is represented by relatively coarse-resolution soil data and finer-resolution terrain and geophysical data (Table 1). We also include spatial measures of soil attribute interpretation reliability associated with the soil variables estimated from the Australian soil classification (McKenzie and Hook 1992, McKenzie *et al*. 2000). We did not consider variables derived from soil water and nutrient balance models, even though these are more proximal, because the coarse-resolution soil data inaccurately partition water availability into soil moisture, runoff and evaporative demand. Instead, we treated the soil and climate variables as independent proxy variables.

We grouped these variables into direct or indirect predictors of plant species distributions and noted whether these were proximal or distal to physiological processes of vegetation growth and development. We also noted the relationship between pairs of variables and judged whether some variables are substitutes for one another and therefore not sensibly combined in the same model. Finally, we ranked each variable according to its perceived relative importance based on ecological rationale and to indicate the order in which the variables might best be combined in a predictive model.

3.3. Correlation and scatter plots

Bivariate scatter plots of variables are a simple yet effective way to view the relationships between pairs of variables and to identify potential outliers and data errors in biological records. We generated scatter plots via the Atlas of Living Australia (http://spatial.ala.org.au).

We introduced a simple alternative to the linear Pearson's correlation coefficient by averaging the range-standardized differences between variables. This amounts to a variant of the Gower Metric (Gower 1971), hereafter referred to as dissimilarity. All values of each layer are first range standardized on a 0:1 ratio scale. The measure then simply takes the average of the differences between each layer for each cell. The resulting dissimilarity values represent a standardized 'volume' between each pair of environmental 'layers', ranging from zero (variables are identical) to one (variables have nothing in common). A truly random relationship produces a Gower dissimilarity value of 0.5. An inverse (complementary) relationship produces a value around 0.7. The continental extent of the variables was used for this analysis. Relationships would change if the comparison was limited to a region or a local area.

3.4. Model building

We developed a repeatable, systematic approach to model building based on a forward stage-wise iterative procedure for testing a large number of correlated variables where it is impractical to test all variables simultaneously. This represents a cautious variant of forward stepwise selection. Our approach has three components. First, relatively independent subgroups of correlated variables are tested using backward elimination with a conservative stopping criterion. Variables dropped in these initial tests are not included in subsequent tests. Second, the remaining candidate variables in each subgroup were

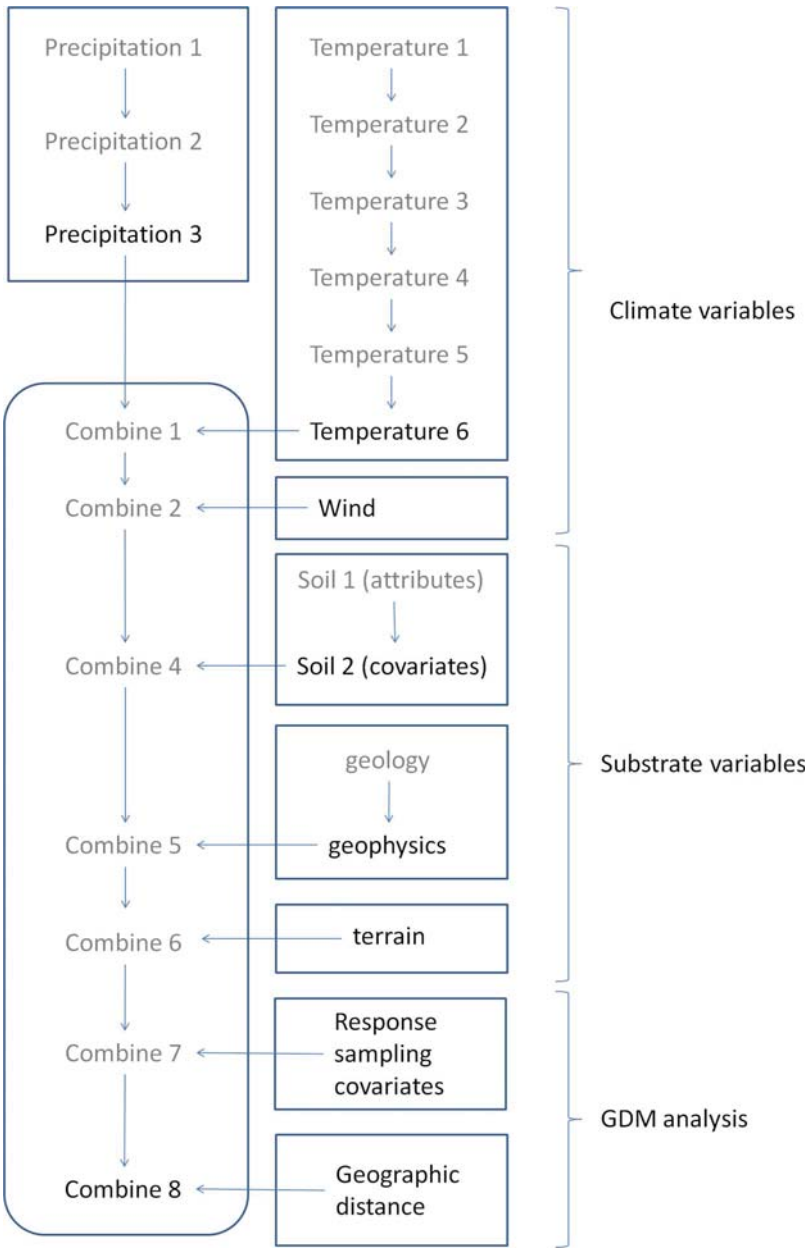

Figure 3. Order in which groups of climate (precipitation, temperature, wind) and substrate (soil, geoscience, terrain) variables were tested and combined in the MaxEnt and GDM models. Groups are defined in the 'test order' column of Table 2 for climate variables and Table 3 for substrate variables. Other variables specific to the GDM model were also tested.

included stage-wise in the MaxEnt and GDM models in order of their group's *a priori* perceived relative importance (Figure 3). As each group of variables was added, a secondary process of backward elimination removed the most marginal variables based on a conservative stopping criterion (specified below). After all candidate variables have been

assessed, the threshold for backward elimination may be increased to find an appropriate trade-off between minimising the number of included variables in the model and excluding variables that explain significant levels of residual variation.

3.4.1. MaxEnt model

Species distribution models derived using ecological knowledge can provide useful information about potential eco-physiological limits and can also be used to infer parameters for more mechanistic models (Williams *et al*. 2000) or experimentation (Battaglia and Williams 1996). We demonstrate this for *E. delegatensis*, a tree species that occurs at higher elevation cool climates in south-eastern Australia, with 216 presence-only records sourced from the Atlas of Living Australia. Records were accepted where spatial precision was less than or equal to 2 km, consistent with the average resolution of the environmental data. For a high proportion of the records, this was less than or equal to 1 km. The analysis region for the background data aims to encompass the expected predicted range of the species (Elith *et al*. 2011). Assuming the presence-only records represent the complete geographic range of the species, the analysis region was defined by the outer envelope of the recorded presences buffered by 0.1° (∼10 km). Curvilinear transformations of the environmental variables were tested using linear, quadratic and product features for each variable. These are preset functions in the ALA implementation of MaxEnt. Variables contributing the least information on the basis of their permutation importance (less than 1% via jack-knife tests) were successively dropped.

3.4.2. GDM model

GDM analyses information about multiple species to predict patterns of beta diversity as a measure of biodiversity for regional assessment and planning. We used the Czekanowski index, also known as the Sørenson index (Czekanowski 1913, Ferrier *et al*. 2007), to measure pair-wise vascular plant species compositional dissimilarity for over 12,000 species representing 83 families (as at September 2009). The response data comprised 875,639 site-pairs, representing a subset of all possible site-pairs randomly selected from strata defined by biogeographic regions (within and between weighted by the number of species), sampling over 100,000 sites (Williams *et al*. 2010). Models were developed using GDM software (Manion 2009a). Curvilinear transformations of the environmental variables were tested using three *I*-spline basis functions (Manion 2009b), defined by the data distribution: minimum, median and maximum percentiles. Significant contributors to spatial patterns of biotic dissimilarity were retained if their partial contribution to the percent deviance explained, in the presence of other variables, was initially greater than 0.01 in variable subgroups, and increased to 0.02 after testing all candidate groups of variables.

3.4.3. Environmental classification

Environmental domain classification has been used as a surrogate for biodiversity patterns to select representative areas for reserves (Mackey *et al*. 1988, Belbin 1995). However, the selection of appropriate environmental variables in a classification often relies on expert opinion (Williams *et al*. 2012). The value of an environmental classification can be maximised by using the variables shown to be important and relevant to the purpose in associated modelling studies. To demonstrate this approach, we applied the subset of variables selected for the vascular plant GDM model to generate a 200 group environmental classification using the ALOC ("Allocation") non-hierarchical clustering algorithm

(Belbin 1987, 1993; http://www.patn.com.au) via the spatial portal of the Atlas of Living Australia (http://spatial.ala.org.au/).

4. Results

4.1. Rationale for environmental variables

Sixty-four environmental variables were evaluated (see framework in Figure 1): 14 water, 13 temperature, 2 radiation, 4 humidity and 4 wind (37 climate); 12 soil, 6 geoscience and 9 terrain (27 substrate) (Table 1). Metadata describing the derivation of these variables can be viewed online from http://spatial.ala.org.au/layers. We classified most of the climate variables as level 2 direct gradients (sensu Guisan and Zimmermann (2000); Figure 2) that are proximally related to the processes of plant growth or development (Tables 2 and 3). For example, the seasonal minimum and maximum rates of change in day and night temperatures are considered to be proximally related to plant phenology responses such as acclimation in photosynthesis during the spring growth flush and frost hardening during autumn.

Examination of scatter plots highlighted an outlier location for *E. delegatensis* at unusually low elevation that was found to be an error. Variables with relatively low dissimilarity[1] (<0.15) showed some close relationships between independently derived climate and substrate variables that are related through a physical environmental process (Tables 2 and 3). For example, the hydrological scoring of pedality based on rainfall infiltration rates (Lin *et al*. 1999) was found to have a dissimilarity value of 0.045 with the minimum monthly rainfall, consistent with the influence of soil moisture on soil morphology (Lin *et al*. 1999). The terrain attributes for slope, relief and roughness, which influence water runoff, were also found to be closely related to water variables, particularly the aridity indices (e.g. ARID_MIN dissimilarity $\sim$0.02). These relationships and the knowledge of how these variables were derived[2] to assist in understanding why some variables were included and others excluded during the iterative model building process, but did not predetermine which variables to test.

We identified five proxy subsets, each containing alternative sets of variables that we judged would not be sensible to combine in the same model (Supplementary Table 1 available online). We also identified variables that were functionally similar but not directly substitutable (Supplementary Table 2). For example, we preferred to not include the precipitation deficit in the same model with rainfall and evaporation (although compare dissimilarity values in Table 2). However, the aridity index being the ratio of rainfall and evaporation, we judged to contain additional information that could be usefully combined with either the precipitation deficit or rainfall and evaporation. For similar reasons, soil depth and clay percentage and its derivative, soil water-holding capacity (Western and McKenzie 2004), were tested separately. Soil water-holding capacity was found to be less effective as a predictor than including both soil depth and percent clay for the GDM models of vascular plant compositional turnover (Supplementary Table 8), but the reverse was marginally the case for the MaxEnt models of *E. delegatensis* (Supplementary Table 5). The dissimilarity values for soil depth and soil water-holding capacity indicate a close relationship (0.08), but both are relatively independent of clay content (Table 3). Clay content in soils influences a wide range of physical processes related to hydrology, mineral status and fertility (Viscarra Rossel 2011). Because alternative sets of water and soil variables (Supplementary Table 1) can influence the selection of other variables included in a model, four different MaxEnt (Section 4.3) and two GDM (Section 4.4) models were compared.

Table 2. Plant ecological rationale for the climate variables.

Variable	Ecological type	Physiological proximity	Plant ecological rationale	Low Dissimilarity (<= 0.15)	Test rank order (T = temperature, P = precipitation, C = combined climate)
MAXTX	Direct, level 2	Proximal	Maximum temperature conditions are generally recorded during the day (after midday) and relate to physiological conditions supporting photosynthesis, respiration and transpiration (water and nutrient uptake) processes, influencing growth and development. The maximum is indicative of the warmest summer conditions that may require specialised drought and heat adaptive responses or morphology.	MINTX (0.0624), TRNGI (0.0871), TRNGA (0.0934), RADNX (0.0937), VPD2MAX (0.1088), MAXTI (0.1381), RTXMAX (0.1395), TMAXABSX (0.1425)	T1 and C1
MAXTI	Direct, level 2	Proximal	Maximum temperature conditions are generally recorded during the day (after midday) and relate to physiological conditions supporting photosynthesis, respiration and transpiration (water and nutrient uptake) processes, influencing growth and development. The coldest maximum is indicative of the winter daytime conditions that may limit capacity for photosynthesis and require avoidance or tolerance/acclimation responses.	RADNI (0.0744), MINTX (0.0938), TRNGX (0.1287), MAXTX (0.1381), TMAXABSX (0.1498)	T1 and C1
MINTI	Direct, level 2	Proximal	Minimum temperature conditions are generally recorded overnight (near dawn) and mainly relate to physiological conditions supporting growth and development or constraints. The minimum is indicative of the depth of winter cold conditions that may require specialised adaptive responses.	TMINABSI (0.0920), GRAVITY (0.1405), EVAPI (0.1444), NUTRIENTS (0.1496)	T1 and C1

MINTX	Direct, level 2	Proximal	Minimum temperature conditions are generally recorded overnight (near dawn) and relate to physiological conditions supporting growth and development or constraints. The warmest minimum is indicative of summer conditions when processes of respiration and metabolism are most active in plants.	MAXTI (0.0938), TRNGA (0.1002), TRNGI (0.1289), RADNI (0.1337), RADNX (0.1399), TRNGX (0.1372)	T1 and C1
RADNX	Direct, level 2	Proximal	Photosynthetic active radiation components are relevant to primary productivity; other wavelength (long and short) influence phenology and behaviour, radiation also drives temperature conditions and is influenced by air humidity and cloud cover (moisture conditions); this calculation is not adjusted by land surface conditions (topographic effects), maximum indicative of high levels of light, correlated with evaporation and temperature.	TMAXABSX (0.0880), TRNGI (0.0923), MAXTX (0.0937), TRNGX (0.1081), TRNGA (0.1163), RTXMAX (0.1174), EVAPX (0.1381), MINTX (0.1399)	T1 and C1
RADNI	Direct, level 2	Proximal	Photosynthetic active radiation components are relevant to primary productivity; other wavelength (long and short) influence phenology and behaviour, radiation also drives temperature conditions and is influenced by air humidity and cloud cover (moisture conditions); this calculation is not adjusted by land surface conditions (topographic effects), minimum indicative of low levels of light, correlated with rainfall and temperature.	MAXTI (0.0744), EVAPI (0.1137), MINTX (0.1337), RTIMAX (0.1341)	T1 and C1
TRNGA	Indirect	Distal	Indicative of annual extremes in temperature range (difference between summer maximum and winter minimum temperatures) and mesoclimatic regime, representing seasonal variation.	RTXMAX (0.0823), TRNGI (0.0894), MAXTX (0.0934), MINTX (0.1002), TRNGX (0.1114), RADNX (0.1163), VPD2MAX (0.1183), TMAXABSX (0.1266), EVAPX (0.1310)	T6 and C1

(*Continued*)

Table 2. (*Continued*).

Variable	Ecological type	Physiological proximity	Plant ecological rationale	Low Dissimilarity (<= 0.15)	Test rank order (T = temperature, P = precipitation, C = combined climate)
TRNGX	Direct, level 2	Proximal	Indicative of daily temperature range (difference between maximum and minimum temperatures) as the physiological response environment, maximum indicates high variation in temperature conditions (possibly inland or continental locations).	TMAXABSX (0.0874), TRNGI (0.0953), RADNX (0.1081), MAXTX (0.1088), RTXMAX (0.1114), TRNGA (0.1114), MAXTI (0.1287), MINTX (0.1372)	T2 and C1
TRNGI	Direct, level 2	Proximal	Indicative of daily temperature range (difference between maximum and minimum temperatures) as the physiological response environment, minimum indicates generally consistent temperature conditions (possibly coastal locations).	MAXTX (0.0871), TRNGA (0.0894), RADNX (0.0923), TRNGX (0.0953), TMAXASBX (0.1084), RTXMAX (0.1097), MINTX (0.1289)	T2 and C1
RTXMAX	Direct, level 2	Proximal	An indicator of phenology shifts due to changes in temperature during periods of significant seasonal shift, such as the onset of spring or the onset of autumn, trends in environmental temperature acts as a cue for behaviour such as the onset of flowering or acclimation conditions such as optimising photosynthesis, maximum temperature conditions occur during daylight hours. Maximum rates of change (rapid rise) in daytime temperatures indicate the onset of spring growing conditions and hardening for summer droughts and herbivore activity.	TRNGA (0.0823), TRNGI (0.1097), TRNGX (0.1114), TMAXABSX (0.1115), RADNX, 0.1174), EVAPX (0.1178), VPD2MAX (0.1273), MAXTX (0.1395), RTIMAX (0.1453)	T3 and C1

RTXMIN	Direct, level 2	Proximal	An indicator of phenology shifts due to changes in temperature during periods of significant seasonal shift, such as the onset of spring or the onset of autumn, trends in environmental temperature act as a cue for behaviour such as the onset of flowering or acclimation conditions such as optimising photosynthesis and maximum temperature conditions occur during daylight hours. Minimum rates of change (rapid fall) in daytime temperatures indicate the onset of autumn and the end of rapid growing conditions, potentially hardening for winter frost or photosynthetic acclimation.	RTIMIN (0.1275), ADEFX (0.1380)	T3 and C1
RTIMAX	Direct, level 2	Proximal	An indicator of phenology shifts due to changes in temperature during periods of significant seasonal shift, such as the onset of spring or the onset of autumn, trends in environmental temperature act as a cue for behaviour such as the onset of flowering or acclimation conditions such as optimising photosynthesis, minimum temperature conditions occur overnight. Maximum rates of change (rapid rise) in night-time temperatures relate to spring respiration and development, hardening for summer droughts and herbivore activity.	VPD2MAX (0.1206), RADNI (0.1341), RTXMAX (0.1453)	T3 and C1

(*Continued*)

Table 2. (*Continued*).

Variable	Ecological type	Physiological proximity	Plant ecological rationale	Low Dissimilarity (<= 0.15)	Test rank order (T = temperature, P = precipitation, C = combined climate)
RTIMIN	Direct, level 2	Proximal	An indicator of phenology shifts due to changes in temperature during periods of significant seasonal shift, such as the onset of spring or the onset of autumn, trends in environmental temperature act as a cue for behaviour such as the onset of flowering or acclimation conditions such as optimising photosynthesis, minimum temperature conditions occur overnight. Minimum rates of change (rapid fall) in night-time temperatures relate to autumn respiration and development, hardening for winter frosts or drought and protecting primordial buds for rapid response when conditions improve.	RTXMIN (0.1275), ADEFX (0.1286), RTIMAX (0.1453), MAGNETICS (0.1494)	T3 and C1
VPD2MAX	Direct, level 1	Proximal	Relevant to plant transpiration processes as the difference between actual and potential air humidity drives evaporative demand when conditions are between freezing and saturation, maximum indicates potential for drought stress depending also on temperature conditions.	EVAPX (0.1064), TRNGA (0.1183), RTIMAX (0.1206), RTXMAX (0.1273)	T4 and C1
VPD2MIN	Direct, level 1	Proximal	Relevant to plant transpiration processes as the difference between actual and potential air humidity drives evaporative demand when conditions are between freezing and saturation, minimum indicates air is saturated with moisture and potential for precipitation, depending also on temperature conditions.	NONE < 0.15	T4 and C1

RH2MAX	Direct, level 1	Proximal	Relevant to evaporative drying processes as the ratio between actual to potential air humidity, maximum indicates air is saturated with moisture.	NONE < 0.15	T4 and C1
RH2MIN	Direct, level 1	Proximal	Relevant to evaporative drying processes as the ratio between actual to potential air humidity, minimum indicates air is dry.	ADEFX (0.1294), ADEFI (0.1010), RAINI (0.1427), RAINX (0.1448), RPRECMAX (0.1492), SLRAIN1 (0.1358)	T4 and C1
TMAXABSX	Direct, level 2	Proximal	Indicative of interannual extremes in temperature, maximum related to periodic extreme heat stress.	TRNGX (0.0874), RADNX (0.0880), TRNGI (0.1084), RTXMAX (0.1115), EVAPX (0.1169), RTNGA (0.1266), MAXTX (0.1425), MAXTI (0.1498)	T5 and C1
TMINABSI	Direct, level 2	Proximal	Indicative of interannual extremes in temperature, minimum related to periodic extreme cold stress.	MINTI (0.0920), GRAVITY (0.1040), EVAPI (0.1404)	T5 and C1
RAINI	Direct, level 2	Proximal	Related to soil water availability and atmospheric moisture, minimum indicates potential drought conditions correlated with low cloud cover and high levels of radiation (rainfall and evaporation together can substitute a measure of precipitation deficit).	ARID_MAX (0.0230), HPEDALITY (0.0449), SLOPE (0.0492), RELIEF (0.0498), ARID_MIN (0.0510), ROUGHNESS (0.0588), ADEFI (0.0634), RAINX (0.0703), RPRECMAX (0.0778), COARSE (0.0857), SLRAIN1 (0.1154), CALCRETE (0.1318), WR_UNR (0.1418), RH2MIN (0.1427), RIDGETOPFLAT (0.1466)	P1 and C1

(*Continued*)

Table 2. (*Continued*).

Variable	Ecological type	Physiological proximity	Plant ecological rationale	Low Dissimilarity (<= 0.15)	Test rank order (T = temperature, P = precipitation, C = combined climate)
RAINX	Direct, level 2	Proximal	Related to soil water availability and atmospheric moisture, maximum indicates potential flood conditions correlated with high cloud cover and low levels of radiation (rainfall and evaporation together can substitute a measure of precipitation deficit).	RPRECMAX (0.0143), ARID_MIN (0.0443), HPEDALITY (0.0493), RELIEF (0.0502), SLOPE (0.0506), ARID_MAX (0.0568), ROUGHNESS (0.0575), ADEFI (0.0583), RAINI (0.0703), COARSE (0.0761), SLRAIN1 (0.0905), RIDGETOPFLAT (0.1426), CALCRETE (0.1446), RH2MIN (0.1448)	P1 and C1
EVAPX	Direct, level 2	Proximal	Drives evapotranspiration, influences soil water availability and atmospheric moisture, maximum indicates potential heat or drought conditions correlated with temperature and radiation (evaporation and rainfall together can substitute a measure of precipitation deficit).	VPD2MAX (0.1064), TMAXABSX (0.1169), RTXMAX (0.1178), TRNGA (0.1310), RADNX (0.1381)	P1 and C1
EVAPI	Direct, level 2	Proximal	Drives evapotranspiration, influences soil water availability and atmospheric moisture, minimum indicates potential flood or cold conditions correlated with temperature and radiation (evaporation and rainfall together can substitute a measure of precipitation deficit).	RADNI (0.1137), MINTI (0.1444), TMINABSI (0.1404)	P1 and C1

ADEFI	Direct, level 2	Distal	Indicative of atmospheric moisture conditions (effective water availability or water stress) as the difference between rainfall and evaporation, maximum is dominated by precipitation (soil environmental not considered). Precipitation deficit can substitute precipitation and evaporation.	RAINX (0.0583), RAINI (0.0634), RPRECMAX (0.0675), SLRAIN1 (0.0719), ARIDX (0.0787), ARIDI (0.0915), RH2MIN (0.1010), MAGNETICS (0.1378)	P1 (substitute) and C1
ADEFX	Direct, level 2	Distal	Indicative of atmospheric moisture conditions (effective water availability or water stress) as the difference between rainfall and evaporation, minimum is dominated by evaporation (soil environmental not considered). Precipitation deficit can substitute precipitation and evaporation.	HPEDALITY (0.0700), SLOPE (0.0910), RELIEF (0.0917), ROUGHNESS (0.1066), COARSE (0.1233), RH2MIN (0.1294), RTIMIN (0.1236), RTXMIN (0.1380)	P1 (substitute) and C1
ARID_MIN	Direct, level 2	Distal	Indicative of atmospheric moisture conditions (relative water availability or water stress) as the ratio between rainfall and evaporation, maximum is dominated by precipitation (soil hydrology attributes not considered).	ROUGHNESS (0.0171), RELIEF (0.0212), SLOPE (0.0222), ARID_MAX (0.0317), HPEDALITY (0.0395), RAINX (0.0443), COARSE (0.0454), RAINI (0.0510), ADEFI (0.0915), CALCRETE (0.1072), RIDEGTOPFLAT (0.1146), SLRAIN1 (0.1293), WR_UNR (0.1365)	P1 (substitute) and C1
ARID_MAX	Direct, level 2	Distal	Indicative of atmospheric moisture conditions (relative water availability or water stress) as the ratio between rainfall and evaporation, minimum is dominated by evaporation (soil hydrology attributes not considered).	RAINI (0.0230), ARIDI (0.0317), SLOPE (0.0319), ROUGHNESS (0.0373), HPEDALITY (0.0401), RAINX (0.0568), RPRECMAX (0.0639), COARSE (0.0659), ADEFI (0.0787), CALCRETE (0.1204), SLRAIN1 (0.1235), WR_UNR (0.1365)	P1 (substitute) and C1

(*Continued*)

Table 2. (*Continued*).

Variable	Ecological type	Physiological proximity	Plant ecological rationale	Low Dissimilarity (<= 0.15)	Test rank order (T = temperature, P = precipitation, C = combined climate)
RPRECMAX	Direct, level 2	Proximal	An indicator of phenology shifts due to changes in rainfall during periods of significant seasonal shift, such as the onset of the tropical wet season in early summer or the onset of the dry season in late autumn, or the change from winter to summer rainfall patterns in Mediterranean regions, maximum indicates increasing rainfall trends.	RAINX (0.0143), ARIDI (0.0457), HPEDALITY (0.0506), RELIEF (0.0522), SLOPE (0.0529), ROUGHNESS (0.0555), ARIDX (0.0639), ADEFI (0.0675), COARSE (0.0735), RAINI (0.0778), SLRAIN1 (0.0922), RIDGETOPFLAT (0.1401), CALCRETE (0.1416), RH2MIN (0.1492)	P2 and C1
RPRECMIN	Direct, level 2	Proximal	An indicator of phenology shifts due to changes in rainfall during periods of significant seasonal shift, such as the onset of the tropical wet season in early summer or the onset of the dry season in late autumn, or the change from winter to summer rainfall patterns in Mediterranean regions, minimum indicates declining rainfall trends.	SLRAIN2 (0.0878)	P2 and C1
SRAIN2MP	Indirect	Distal	Ratio of annual contrast in regional rainfall conditions that may relate to phenology shifts between spring and autumn equinox conditions (for continental scale assessments substituted by slrain2).	RH2MAX (0.1417)	P3 (substitute) and C1

SRAIN1MP	Indirect	Distal	Ratio of annual contrast in regional rainfall conditions that may relate to phenology shifts between summer and winter solstice conditions (for continental scale assessments substituted by slrain1).	ROUGHNESS (0.0220), ARID_MIN (0.0250), SLOPE (0.0357), RPRECMAX (0.0419), RAINX (0.0425), COARSE (0.0443), HEDALITY (0.0499), ARIDX (0.0530), RAINI (0.0729), ADEFI (0.0920), CALCRETE (0.1109), RIDGETOPFLAT (0.1149), SLRAIN1 (0.1223), WR_UNR (0.1462)	P3 (substitute) and C1
SLRAIN2	Indirect	Distal	A factor ratio of annual rainfall contrasts for continental scale assessment of rainfall seasonality (equinox comparison of conditions), particularly distinguishes wet tropical from Mediterranean rainfall regimes (for regional scale assessments substituted by srain2mp).	RPRECMIN (0.0878)	P3 and C1
SLRAIN1	Indirect	Distal	A factor ratio of annual rainfall contrasts for continental scale assessment of rainfall seasonality (solstice comparison of conditions), particularly distinguishes wet tropical from Mediterranean rainfall regimes (for regional scale assessments substituted by srain1mp).	ADEFI (0.0719), RAINX (0.0905), RPRECMAX (0.0922), HPEDALITY (0.0988), RAINI (0.1154), ARID_MAX (0.1235), ARID_MIN (0.1293), RELIEF (0.1294), SLOPE (0.1294), RH2MIN (0.1358), COARSE (0.1453), ROUGHNESS (0.1402)	P3 and C1
WINDRX	Direct, level 2	Distal	Indicative of wind disturbance regime.	WINDRI (0.0425), WINDSPMAX (0.0506), WINDSPMIN (0.0870), MAGNETICS (0.1078)	C2
WINDRI	Direct, level 2	Distal	Indicative of wind disturbance regime.	WINDRX (0.0425), WINDSPMAX (0.0461), WINDSPMIN (0.0659), MAGNETICS (0.0920)	C2

(*Continued*)

Table 2. (*Continued*).

Variable	Ecological type	Physiological proximity	Plant ecological rationale	Low Dissimilarity (<= 0.15)	Test rank order (T = temperature, P = precipitation, C = combined climate)
WINDSPMAX	Direct, level 2	Distal	Indicative of maximum wind regime, potentially damaging.	WINDRI (0.0461), WINDRX (0.0506), WINDSPMIN (0.0750), MAGNETICS (0.0802)	C2
WINDSPMIN	Direct, level 2	Distal	Indicative of minimum wind regime.	WINDRI (0.0659), WINDSPMAX (0.0750), WINDRX (0.0870), MAGNETICS (0.0895)	C2

Notes: Test order defines the *a priori* order in which variable sets are included in a model and tested for significance (as shown in Figure 3). Climate variables are derivatives of the monthly variables generated from ANUCLIM 5.1 (Hutchinson *et al*. 2000). Metadata for each variable can be viewed from the links provided at http://spatial.ala.org.au/layers. Dissimilarity is from the association matrix (http://spatial.ala.org.au/files/inter_layer_association.csv). The full metadata can be viewed from the link for each variable provided at http://spatial.ala.org.au/layers.

Table 3. Plant ecological rationale for the substrate variables.

Variable	Ecological type	Physiological proximity	Plant ecological rationale	Low Dissimilarity (<= 0.15)	Test order (S = soil, G = geoscience, T = terrain)
NUTRIENTS	Direct, level 2	Proximal	Generally indicative of potential soil nutrient status (essential mineral resources).	CLAY (0.1417), MAGNETICS (0.1099), SOLDEPTH (0.1312), TWI (0.1333), MINTI (0.1496)	S1 and C4
SOLDEPTH	Direct, level 2	Proximal	Soil depth can affect root exploration volume for water and nutrients as well as water-holding capacity (with clay, acts as an alternative to SOLPAWHC).	MAGNETIC (0.1492), NUTRIENTS (0.1312), SOLPAWHC (0.0785)	S1 and C4
CLAY	Direct, level 2	Proximal	Affects the nutrient and soil moisture capacity as well as pedal structure which also affects root exploration capacity and soil drainage (with soil depth, acts as an alternative to SOLPAWHC).	NUTRIENTS (0.1417)	S1 (substitute for clay, soldepth) and C4
SOLPAWHC	Direct, level 2	Proximal	Combination of soil texture and soil depth indicates soil moisture potential within plant-available range (field capacity to wilting point) (alternative to soil depth and clay percent).	SOLDEPTH (0.0785), MAGNETIC (0.1312), MINTI (0.1479)	S1 and C4
CALCRETE	Direct, level 2	Distal	Indirectly related to soil depth and water-holding capacity and root exploration difficulty.	ROUGHNESS (0.0937), RELIEF (0.1106), SRAIN1MP (0.1109), SLOPE (0.1122), COARSE (0.1177), HPEDALITY (0.1198), RAINI (0.1318), SPRECMAX (0.1416), RAINX (0.1446)	S1 and C4
KSAT	Direct, level 2	Distal	Relevant to soil water balance in determining how water moves through the soil and deep drainage.	NONE < 0.15	S1 and C4

(*Continued*)

Table 3. (*Continued*).

Variable	Ecological type	Physiological proximity	Plant ecological rationale	Low Dissimilarity (<= 0.15)	Test order (S = soil, G = geoscience, T = terrain)
HPEDALITY	Direct, level 2	Distal	Relevant to soil physical structure which relates to the ability of roots to explore the soil and moisture potential, ordered by hydrological capacity.	ADEFX (0.0700), ARID_MIN (0.0395), ARID_MAX (0.0401), RAINI (0.0449), RAINX (0.0493), RPRECMAX (0.0506), SLRAIN1 (0.0988), SRAIN1MP (0.0499), CALCRETE (0.1198), RIDGETOPFLAT (0.1322)	S1 and C4
BDENSITY	Direct, level 2	Distal	Related to harshness of the soil environment for root exploration and water infiltration capacity.	NONE < 0.15	S1 and C4
COARSE	Direct, level 2	Distal	Describes the stoniness of the soil environment which affects root exploration and hydrology (drainage).	ROUGHNESS (0.0330), SRAIN1MP (0.0443), ARID_MIN (0.0454), ARID_MAX (0.0659), SLOPE (0.0516), HPEDALITY (0.0699), RAINI (0.0857), RAINX (0.0761), RPRECMAX (0.0735), CALCRETE (0.1177), RIDGETOPFLAT (0.1132), ADEFX (0.1233), SLRAIN1 (0.1453)	S1 and C4
DATASUPT	Indirect	Covariate	General uncertainty covariate for interpretation of soil attributes from soil profile form.	NONE < 0.15	S2 and C4
KS_ERR	Indirect	Covariate	Uncertainty covariate for corresponding soil attribute variable (KSAT).	MAGNETIC (0.1085)	S2 and C4

WR_UNR	Direct, level 2	Distal	Uncertainty covariate for corresponding soil attribute variable (SOLPAWHC).	ROUGHNESS (0.1330), ARID_MIN (0.1365), ARID_MAX (0.1365), HPEDALITY (0.1377), RELIEF (0.1363), SLOPE (0.1368), RAINI (0.1418), SRAIN1MP (0.1462)	S2 and C4
FERT	Direct, level 2	Proximal	Indicative of potential fertility of weathered rock material based on an interpretation of surface geology.	NONE < 0.15	G1 and C5
GEOLLRNGEAGE	Indirect	Distal	Rock age is indirectly related to substrate weathering and soil formation and nutrient status (range indicates heterogeneity in ages in a map unit).	GEOLLRNGEAGE (0.0708)	G1 and C5
GEOLLMEANAGE	Indirect	Distal	Rock age is indirectly related to substrate weathering and soil formation and nutrient status (mean age).	GEOLLMNAGE (0.0708)	G1 and C5
GRAVITY	Indirect	Distal	Indirectly related to vegetation through surface rock type, gravity anomalies are caused by lateral density contrasts within the sedimentary section, crust and subcrust of the earth. Gravity highs reflect areas of crustal thinning and gravity lows indicate areas of low-density sediment infill (Gunn 1997). Mafic ore bodies may cause regional magnetic anomalies and add to gravity highs.	Nil substrate < 0.15, MINTI (0.1405), TMINABSI (0.1040)	G2 and C5
MAGNETICS	Indirect	Distal	Indirectly related to vegetation through surface rock type. Magnetic anomalies indicate weathering and sedimentation processes and strata type including hydrological features such as ground water supply. Natural remnant magnetism of a rock reflects its history and the ambient field including some surface exposure to lightening, as well as rock metamorphism and hydrothermal reactions (Clark 1997).	WINDSPMAX (0.0802), WINDSPMIN (0.0895), WINDRI (0.0920), TWI (0.0979), WINDRX (0.1078), KS_ERR (0.1085), NUTRIENTS (0.1099), MINTI (0.1311), SOLPAWHC (0.1312), ADEFI (0.1378), SOLDEPTH (0.1492), RTIMIN (0.1494)	G2 and C5

(*Continued*)

Table 3. (*Continued*).

Variable	Ecological type	Physiological proximity	Plant ecological rationale	Low Dissimilarity (<= 0.15)	Test order (S = soil, G = geoscience, T = terrain)
WII_WGS1KB	Indirect	Distal	Indirectly related to vegetation through soil depth, texture and essential nutrients. The weathering intensity index was developed from regression models for erosional landscapes but has the potential to inform deposition processes and materials (Wilford 2012). As weathering intensity increases there are changes in the hydrological, geochemical and geophysical characteristics of the regolith.	NONE < 0.15	G2 and C5
TWI	Indirect	Distal	Relevant to soil deposition and erosion processes in landscape driven by topography and weathering, and also related to runoff properties and soil moisture.	MAGNETICS (0.0979), NUTRIENTS (0.1333), WINDSPMIN (0.1397), MINTI (0.1421), WINDSPMAX (0.1427)	T1 and C6
EROSIONAL	Indirect	Distal	These are the hillslopes where erosional rather than depositional processes dominate.	NONE < 0.15	T1 and C6
MRRTF	Indirect	Distal	Median value of ridgetop flatness (potentially relates to land stability and substrate conditions).	NONE < 0.15	T1 and C6
RIDGETOPFLAT	Indirect	Distal	Ridgetop flats classified from multi-resolution terrain analysis (potentially relates to land stability and substrate conditions).	ROUGHNESS (0.1040), COARSE (0.1132), ARID_MIN (0.1146), SRAIN1MP (0.1149), RELIEF (0.1200), SLOPE (0.1217), HPEDALITY (0.1322), RAINX (0.1424), RAINI (0.1466)	T1 and C6
RELIEF	Indirect	Distal	Elevation range – potentially related to local terrain complexity and associated substrate conditions.	SLOPE (0.0038), ARID_MIN (0.0212), ROUGHNESS (0.0220), HPEDALITY (0.0439), RAINI (0.0498), COARSE (0.0499), RAINX (0.0502), ADEFX (0.0917), CALCRETE (0.1106), RIDGETOPFLAT (0.1200), SLRAIN1 (0.1294), WR_UNR (0.1363)	T1 and C6

ROUGHNESS	Indirect	Distal	Coefficient of variation in elevation potentially related to local terrain heterogeneity and associated substrate conditions.	ARID_MIN (0.0171), RELIEF (0.0220), SRAIN1MP (0.0220), SLOPE (0.0238), COARSE (0.0330), ARID_MAX (0.0373), HPEDALITY (0.0539), RPRECMAX (0.0555), RAINX (0.0575), RAINI (0.0588), CALCRETE (0.0932), RIDGETOPFLAT (0.1040), ADEFX (0.1066), WR_UNR (0.1330), SLRAIN1 (0.1402)	T1 and C6
SLOPE	Indirect	Distal	Slope potentially relates to land stability and substrate conditions, and to runoff processes and hence soil moisture regime.	RELIEF (0.0038), ARID_MIN (0.0222), ROUGHNESS (0.0238), ARID_MAX (0.0319), SRAIN1MP (0.0357), HPEDALITY (0.0449), RAINI (0.0492), RAINX (0.0506), COARSE (0.0516), RPRECMAX (0.0529), ADEFX (0.0910), CALCRETE (0.1122), RIDGETOPFLAT (0.1217), SLRAIN1 (0.1294), WR_UNR (0.1368)	T1 and C6
VALLEYBOTTOM	Indirect	Distal	Valley bottoms potentially relate to land stability and substrate conditions.	NONE < 0.15	T1 and C6
MRVBF	Indirect	Distal	Median value of valley bottom flatness (potentially relates to land stability and substrate conditions).	NONE < 0.15	T1 and C6

Notes: Test order defines the *a priori* order in which substrate variables are combined after testing the climate variables (as shown in Figure 3). Soil variables are interpreted from the 1:1M-1:3M Atlas of Australian Soils. Geology variables are interpreted from the 1:1M Geology Atlas. Terrain attributes are derived from the 9 seconds DEM of Australia. Metadata for each variable can be viewed from the links provided at http://spatial.ala.org.au/layers. Dissimilarity is from the association matrix (http://spatial.ala.org.au/files/inter_layer_association.csv).

4.2. Variable selection through model building

Variables were grouped into subsets for successive testing and combined in order of their perceived importance to the model based on prior experience in continental modelling (Figure 3): water, energy (mainly temperature), soil, geology and terrain (test order in Tables 2 and 3). Potentially redundant, correlated variables in each group were identified and removed via backward elimination, using the respective thresholds for MaxEnt and GDM, before stage-wise combining with other groups of variables (Figure 3). These thresholds are specific to the model and were judged reasonable trade-offs between greater parsimony and reduced explanation.

4.3. MaxEnt model of E. delegatensis

Four alternative MaxEnt models were derived using substitutable subsets of environmental variables. These models were approximately equivalent in terms of overall performance and prediction with 13 of the 24 variables in common, albeit with different relative contribution and permutation importance (see Supplementary Tables 3–5). Of the 64 variables tested, 32 were included in at least one of these models (Table 4 and Supplementary Table 9). Across all four models, the most important variable was the hottest monthly maximum temperature (MAXTI) (Supplementary Table 3). Climatic energy variables (temperature, radiation and humidity) contributed strongly to the scaling of the model (65–70%), followed by water (12–18%). Substrate variables (∼16%) were approximately equivalent in importance across the soil, geosciences and terrain subgroups (Supplementary Table 4). Based on permutation importance however, the contrasts between energy and water groups declined (Supplementary Table 5). In some cases, terrain variables were approximately equal or more important than energy variables. Terrain is locally important to climate and soil formation throughout the range of this species in montane regions of south-eastern Australia. Prediction maps derived from the four models are reasonably consistent (Supplementary Figure 1). The mean and standard deviation of the four models provide a visual depiction of this consistency (Figure 4).

4.4. GDM model of vascular plants compositional turnover

Two models of vascular plant compositional turnover were derived using substitutable subsets of environmental variables: either precipitation and evaporation or precipitation deficit (Supplementary Tables 6–8). These models explained about 50% of the deviance

Table 4. Numbers of environmental variables used in four MaxEnt and two GDM models in the broad groups defined in Table 1 (details in Supplementary Table 9).

Group[a]	No. of variables	MaxEnt	GDM	Total used
Water	14	9	8	11
Energy	23	9	9	13
Soil	12	6	4	9
Geoscience[b]	6	3	4	5
Terrain	9	5	3	6
Total	64	32	28	44

Notes: [a]Group follows Table 1.
[b]Geoscience refers to geology and geophysical variables.

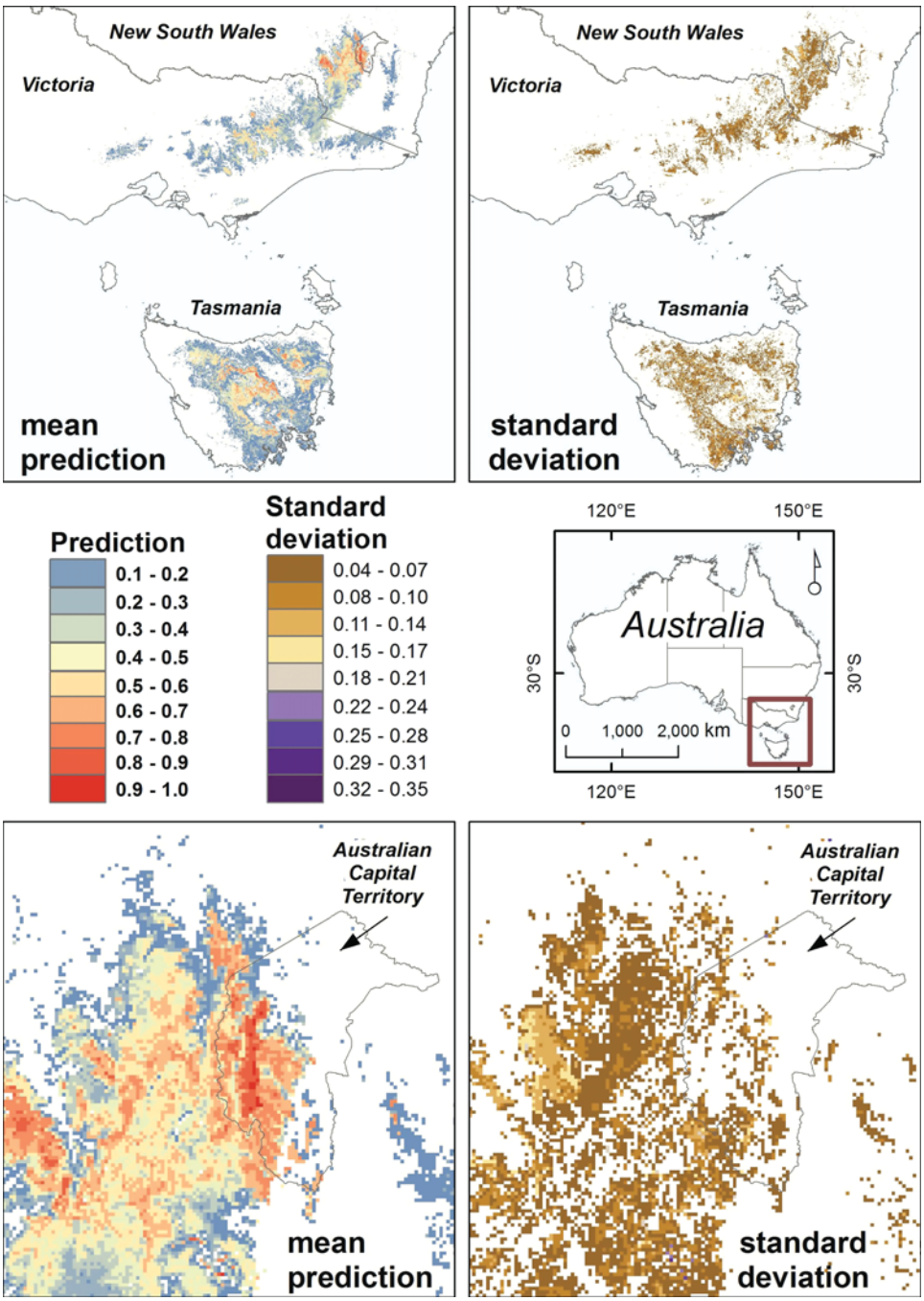

Figure 4. Ensemble mean probability of presence (>0.1) and standard deviation (>0.04) for the natural distribution of *Eucalyptus delegatensis* based on four alternative MaxEnt models shown for the analysis extent in south eastern Australia (top) and in detail for the region focussed on the Australian Capital Territory (bottom). White areas for the mean prediction indicate values <0.1 and for the standard deviation, values <0.04.

and included up to 28 predictors (Supplementary Table 6). The majority of variables are in common, varying slightly in their relative contribution to turnover (summed predictor coefficient values), and with greater contrasts in their partial deviance explained (Supplementary Tables 7 and 8). Water variables contributed slightly more than temperature variables (37–43% vs. 29–34%), followed in order by geophysical, soil and terrain variables (Supplementary Table 7).

Additional covariates for sampling adequacy and geographic distance between locations were included in the GDM analysis (Figure 3). The sampling covariates, which take into account sampling inadequacies through the number of species and observation records aggregated at the 0.01° grid scale (Williams *et al*. 2010), reduced the model intercept and slightly influenced other predictor coefficients (Supplementary Table 8). However, geographic distance significantly influenced other predictor coefficients (decreases) but increased the model intercept (Supplementary Table 8). The inclusion of geographic distance contributed additional information about the response potentially related to latent variables, evolutionary history, dispersal barriers and non-equilibrium conditions. Burley *et al*. (2012) apply more rigorous tests in evaluating the spatial structure of biodiversity–environment turnover relationships using GDM.

4.5. Environmental classification

The ALOC 200 group environmental domain classification in Figure 5 is based on the variables used in a GDM model (Model 2, Supplementary Table 8) and coloured according to similarity in group relationships using the full colour spectrum (Belbin *et al*. 1983). The classification distinguishes the environments of eastern Australia (red, mauve and brown hues) from those of central and western Australia (blue, green and yellow hues). Colour trends are comparable with the agro-climatic regions defined by Hutchinson *et al*. (2005). Although the ecological model weights and transforms each variable according to

Figure 5. An example 200 group environmental domain classification (ALOC algorithm) of the environmental variables selected and scaled by GDM Model 2 (see Supplementary Information). Similar colours represent similarity in group relationships (Belbin *et al*. 1983).

the relationship with the response in the presence of other variables in the model, the classification here weights each variable equally. Potential approaches to incorporating non-equal weighting of variables in this type of classification include using transformations of variables generated by a GDM model (Ferrier *et al*. 2007) or assigning weights according to expert opinion (Williams *et al*. 2012).

5. Discussion

5.1. The Data model

This work was motivated by the need to provide guidance on which environmental variables to include in a biodiversity model, given a large number of candidate variables. Although the data model concept was developed for species distribution modelling (Austin 2002b, 2007, Franklin 2009), it is also applicable to macro-ecological modelling of species assemblages using methods such as GDM (Ferrier *et al*. 2007) and to environmental classifications/domains for biodiversity assessment (Belbin 1987, 1993, Mackey *et al*. 2008). Our Australian case studies illustrate how to address the general problem of choosing environmental variables appropriate to a study. Selected variables can be used in applications requiring a biodiversity surrogate, such as environmental classification (Figure 5) and analysis of biological survey gaps (Funk *et al*. 2005). The approach uses applicable ecological theory to guide the compilation of environmental information followed by numerical evaluation of relationships examined in the light of ecological knowledge (Figure 1). Through an expansion of the data model (sensu Austin 2002b), issues of data accuracy, precision, resolution, extent and fitness for purpose can be systematically addressed.

5.2. Pair-wise dissimilarity – a novel measure of variable relationships

A variant of the Gower metric was developed as an improved measure of the relationship between pairs of variables. This measure is used by the Atlas of Living Australia (http://spatial.ala.org.au) to inform users of the relationship between selected and unselected variables based on their continental extents.

The advantage of the dissimilarity measure over the Pearson's correlation coefficient is that non-linear relationships between variables are better accounted for. However, it should be noted that the usefulness of a set of variables selected through correlation analysis alone may not necessarily represent a minimum set describing a particular biological response due to interactions among predictors and their relationship to the response. Further work is needed to clarify the effectiveness of dissimilarity matrices as a tool alongside an ecological rationale for *a priori* selection of environmental variables used in a biodiversity model.

5.3. Approach to variable selection using GDM and MaxEnt

Variable selections in the two case study biodiversity models (GDM and MaxEnt) were generally consistent, but were affected to some extent by the choice of substitutable variables which slightly alter the relationship with other included variables. Of the 64 candidate variables considered, 44 were used in one or more of the four MaxEnt and two GDM models (Table 4). The majority of the water variables were used, about half of the 'energy' variables and a variety of substrate variables (Supplementary Table 9). In each case, the stage-wise inclusion of selected variables, followed by a repeatable process of elimination,

ensured that potentially over-fitted models were successively trimmed back. This approach results in a slightly richer model because it retains important confounding variables that provide a required adjustment for one or more of the variables remaining in the model (Bursac *et al*. 2008).

The threshold used for backward elimination requires a subjective judgement by the analyst balancing parsimony and explanation. We used a threshold minimum of 1% relative permutation importance for a variable to be retained in the MaxEnt analyses and 0.02 minimum for partial percent deviance explained in the GDM analyses. Removal of marginally significant variables moderated the potential to over-fit the models (Supplementary Tables 3 and 6). More stringent criteria can be applied to further limit the number of included variables (Table 4) balanced by the potential to under-fit the model. Our case studies confirm that *a priori* decisions about which variables to include/exclude from a model cannot easily be made without applying a structured testing process.

Overall, a systematic approach to environmental variable evaluation and selection through model building provides a basis for more explicit linking of the results with ecological theory. Tests with different species or assemblages are needed to clarify which of these findings, in terms of variables consistently selected, are generic and which are specific to this case study (Austin 2002a). However, the principles raised here are likely to be of broader relevance.

5.4. Direct gradient analysis

In order to make the best use of the modelling effort, the results should be evaluated for rational patterns both as a geographic map of predictions and as response curves in environmental space via direct gradient analysis (Austin *et al*. 1990). The latter provides opportunities for critique, inference and synthesis in the context of ecological theory. For example, Williams *et al*. (2000) developed a simple graphical method to enable the results of a multivariate predictive model to be visualised along a single environmental gradient and related to the ecological continuum concept (Austin and Smith 1989). These results can be used to propose hypotheses about physiological processes or biotic interactions, determining the ecotone between co-occurring eucalypts (Battaglia and Williams 1996). For example, MaxEnt models of *E. delegatensis* consistently suggest a minimum temperature limit around −5°C possibly due to frost damage, a functional range between 0°C and 25°C and a broad growth optimum around 6–15°C. This temperature optimum reasonably approximates the regeneration niche for the species based on a detailed study of the germination response (Battaglia 1997). Such findings have potential application as hypotheses for niche parameters in a more mechanistic or process-based model of vegetation growth and productivity (Coops *et al*. 2007) and are essential to understanding range-shifts for climate change forecasting (Kearney and Porter 2009, Fordham *et al*. 2011, Pagel and Schurr 2012).

6. Conclusions

Although it is not possible to establish a generic set of predictors that will be applicable in all cases, ecological theory provides the rationale for why particular predictors should be sought and used. Frameworks such as Figure 2, developed by Guisan and Zimmermann (2000), are a first step in identifying environmental variables of relevance in vegetation models. A systematic approach to the selection of environmental variables can be expected to lead to more robust predictive models and engender increased understanding of biodiversity–environment relationships.

Although physiologically based predictors are preferred (e.g. level 1 variables in Guisan and Zimmermann (2000)) (Figure 2), these are rarely available at high enough resolution to be effective in all situations. Indirect factors may be justified where they correlate consistently with direct gradients (Austin 2002b). It is important to state the reasons (hypotheses) for inclusion or exclusion of a predictor on '*a priori*' or other grounds such as error in primary data (Austin and Van Niel 2011), inadequate spatial resolution (Thomas *et al*. 2002) or confounding complexity (Peterson and Nakazawa 2008).

Advances in spatial analysis and remote detection of environment are generating an increasing variety of environmental variables that may describe proximal processes relevant to species distributions and so increase the generality of predictive models. However, a large number of legacy variables coexist with these new variables in environmental databases, requiring a structured process for deciding which variables to use in a biodiversity model. Even in relatively data-poor regions, a wide range of direct and indirect environmental variables can be gathered from global data aggregations (see http://daac.ornl.gov/, http://www.worldclim.org/). For example, elevation from the shuttle radar topography mission (Rabus *et al*. 2003) is the foundation of terrain indices (Wilson and Gallant 2000), and WorldClim (Hijmans *et al*. 2005) provides analogous climate layers to those demonstrated here.

Our case studies suggest some key components of a structured approach to selecting environmental variables that are likely to be of broad relevance for other taxa and regions: (1) a cost-effective compilation of variables in the context of an explicit ecological framework for the study, knowledge of attribute accuracy and resolution; (2) rigorous evaluation of non-linear relationships between variables using process knowledge of variable origin and development, scatter plots and dissimilarity matrices; (3) selection and grouping of variables based on hypotheses of relative ecological importance and perceived predictor effectiveness; (4) systematic testing of variables as predictors through model building and refinement and (5) model critique, inference and synthesis using direct gradient analysis to evaluate the shape of response curves in the context of ecological theory by presenting the prediction in both geographic and environmental space.

Acknowledgements

Financial support was provided through the Atlas of Living Australia, the Australian Government's Caring for our Country initiative and CSIRO Ecosystem Sciences through the Building Resilient Australian Biodiversity Assets science theme. The GDM analysis used biological data compiled in the Australian Natural Heritage Assessment Tool database, accessed with assistance from Dan Rosauer, Tania Laity, Jonathan Face and Anthony Whalen; and software developed by Glenn Manion. The MaxEnt analysis used data aggregated by the Atlas of Living Australia. Part of this work arose through the authors' activities within the Working Group: 'Multidisciplinary approaches to key Australian biodiversity challenges of 2010 and beyond', convened by Daniel P. Faith and Simon Ferrier, within the ARC Research Network for understanding and managing Australian biodiversity (the Environmental Futures Network). The authors acknowledge the work of Adam Collins and Ajay Ranipeta for their development of the environmental library and analytical tools in the Spatial Portal of the Atlas of Living Australia. Randal Storey contributed to the development of the 1 km gridded environmental data. Comments by Margaret Cawsey, Michael Doherty, Janet Franklin and an anonymous reviewer improved the presentation of this article.

Notes

1. The full association matrix can be downloaded from http://spatial.ala.org.au/files/inter_layer_association.csv.
2. The full metadata can be viewed from the link for each variable provided at http://spatial.ala.org.au/layers.

References

Anderson, R.P. and Raza, A., 2010. The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus Nephelomys) in Venezuela. *Journal of Biogeography*, 37 (7), 1378–1393.

Araújo, M.B. and Guisan, A., 2006. Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, 33 (10), 1677–1688.

Ashcroft, M., Chisholm, L., and French, K., 2008. The effect of exposure on landscape scale soil surface temperatures and species distribution models. *Landscape Ecology*, 23 (2), 211–225.

Ashcroft, M.B., French, K.O., and Chisholm, L.A., 2011. An evaluation of environmental factors affecting species distributions. *Ecological Modelling*, 222 (3), 524–531.

Austin, M.P., 1998. An ecological perspective on biodiversity investigations: examples from Australian eucalypt forests. *Annals of the Missouri Botanical Garden*, 85 (1), 2–17.

Austin, M.P., 2002a. Case studies of the use of environmental gradients in vegetation and fauna modeling: theory and practice in Australia and New Zealand. *In*: J.M. Scott *et al*., eds. *Predicting species occurrences: issues of accuracy and scale*. Island Press, Covelo, California. 73–82.

Austin, M.P., 2002b. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, 157 (2–3), 101–118.

Austin, M.P., 2005. Vegetation and environment discontinuities and continuities. *In*: E. Van Der Maarel, ed. *Vegetation ecology*. Oxford: Blackwell Publishing, 52–84.

Austin, M.P., 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling*, 200 (1–2), 1–19.

Austin, M.P., Nicholls, A.O., and Margules, C.R., 1990. Measurement of the realized qualitative niche: environmental niches of 5 Eucalyptus species. *Ecological Monographs*, 60 (2), 161–177.

Austin, M.P., Pausas, J.G., and Nicholls, A.O., 1996. Patterns of tree species richness in relation to environment in southeastern New South Wales, Australia. *Australian Journal of Ecology*, 21 (2), 154–164.

Austin, M.P. and Smith, T.M., 1989. A new model for the continuum concept. *Vegetatio*, 83 (1–2), 35–47.

Austin, M.P. and Van Niel, K.P., 2011. Improving species distribution models for climate change studies: variable selection and scale. *Journal of Biogeography*, 38 (1), 1–8.

Battaglia, M., 1997. Seed germination model for *Eucalyptus delegatensis* provenances germinating under conditions of variable temperature and water potential. *Functional Plant Biology*, 24 (1), 69–79.

Battaglia, M. and Williams, K.J., 1996. Mixed species stands of eucalypts as ecotones on a water supply gradient. *Oecologia*, 108 (3), 518–528.

Belbin, L., 1987. The use of non-hierarchical allocation methods for clustering large sets of data. *The Australian Computer Journal*, 19 (1), 32–41.

Belbin, L., 1993. Environmental representativeness: regional partitioning and reserve selection. *Biological Conservation*, 66 (3), 223–230.

Belbin, L., 1995. A multivariate approach to the selection of biological reserves. *Biodiversity and Conservation*, 4 (9), 951–963.

Belbin, L., Marshall, C., and Faith, D., 1983. Representing relationships by automatic assignment of color. *Australian Computer Journal*, 15 (4), 160–163.

Burley, H., Laffan, S., and Williams, K.J., in press. Spatial non-stationarity and anisotropy of compositional turnover in eastern Australian *Myrtaceae* species. *International Journal of Geographic Information Science*, in press.

Bursac, Z., *et al*., 2008. Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine*, 3 (1), 17.

Clark, D.A., 1997. Magnetic petrophysics and magnetic petrology: aids to geological interpretation of magnetic surveys. *AGSO Journal of Australian Geology and Geophysics*, 17 (2), 83–103.

Coops, N.C., Coggins, S.B., and Kurz, W.A., 2007. Mapping the environmental limitations to growth of coastal Douglas-fir stands on Vancouver Island, British Columbia. *Tree Physiology*, 27 (6), 805–815.

Czekanowski, J., 1913. *Zarys metod statystcznych w zastosowaniu do anthropologii [An outline of statistical methods applied in anthropology]*. Warsaw: Towarzystwo Naukowe Warszawskie.

Elith, J., *et al*., 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29 (2), 129–151.

Elith, J., *et al*., 2011. A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, 17 (1), 43–57.

Elith, J., Kearney, M., and Phillips, S., 2010. The art of modelling range-shifting species. *Methods in Ecology and Evolution*, 1 (4), 330–342.

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

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

Ferrier, S., *et al*., 2007. Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Diversity and Distributions*, 13 (3), 252–264.

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

Fordham, D.A., *et al*., 2011. Plant extinction risk under climate change: are forecast range shifts alone a good indicator of species vulnerability to global warming? *Global Change Biology*, 18 (4), 1357–1371.

Franklin, J., 1995. Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography*, 19 (4), 474–499.

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

Funk, V.A., Richardson, K.S., and Ferrier, S., 2005. Survey-gap analysis in expeditionary research: where do we go from here? *Biological Journal of the Linnean Society*, 85, 549–567.

Gallant, J. and Read, A., 2009. Enhancing the SRTM data for Australia. *In:* R. Purves, *et al*., eds. *Proceedings of geomorphometry 2009*. Zurich: University of Zurich, 149–154.

Gower, J.C., 1971. A general coefficient of similarity and some of its properties. *Biometrics*, 27 (4), 857–871.

Graf, R.F., *et al*., 2005. The importance of spatial scale in habitat models: capercaillie in the Swiss Alps. *Landscape Ecology*, 20 (6), 703–717.

Guisan, A., *et al*., 2007. Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions*, 13 (3), 332–340.

Guisan, A. and Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, 8 (9), 993–1009.

Guisan, A. and Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135 (2–3), 147–186.

Gunn, P.J., 1997. Regional magnetic and gravity responses of extensional sedimentary basins. *AGSO Journal of Australian Geology and Geophysics*, 17, 2.

Hagen-Zanker, A., 2009. An improved Fuzzy Kappa statistic that accounts for spatial autocorrelation. *International Journal of Geographical Information Science*, 23 (1), 61–73.

Harmsen, E.W., *et al*., 2009. Seasonal climate change impacts on evapotranspiration, precipitation deficit and crop yield in Puerto Rico. *Agricultural Water Management*, 96 (7), 1085–1095.

Hastie, T., *et al*., 2007. Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics*, 1, 1–29.

Hijmans, R.J., *et al*., 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25 (15), 1965–1978.

Hosmer, D.W.J. and Lemeshow, S., 1989. *Applied logistic regression*. New York: Wiley.

Hutchinson, M.F., *et al*., 2000. *ANUCLIM users guide version 5.1*. Canberra: Centre for Resource and Environmental Studies, The Australian National University.

Hutchinson, M.F., *et al*., 2005. Integrating a global agro-climatic classification with bioregional boundaries in Australia. *Global Ecology and Biogeography*, 14 (3), 197–212.

Hutchinson, M., *et al*., 2008. *GEODATA 9 second DEM and D8. Digital elevation model version 3 and flow direction grid. Gridded elevation and drainage data. Source scale 1:250 000. User guide*. 3rd ed. Canberra: Fenner School of Environment and Society, The Australian National University and Geoscience Australia, Australian Government.

Jeffrey, S.J., *et al*., 2001. Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environmental Modelling & Software*, 16 (4), 309–330.

Kaliontzopoulou, A., *et al*., 2008. Modelling the partially unknown distribution of wall lizards (Podarcis) in North Africa: ecological affinities, potential areas of occurrence, and methodological constraints. *Canadian Journal of Zoology*, 86 (9), 992–1001.

Kazan, K. and Manners, J.M., 2011. The interplay between light and jasmonate signalling during defence and development. *Journal of Experimental Botany*, 62 (12), 4097–4100.
Kearney, M. and Porter, W., 2009. Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters*, 12 (4), 334–350.
Kent, R. and Carmel, Y., 2011. Presence-only versus presence–absence data in species composition determinant analyses. *Diversity and Distributions*, 17 (3), 474–479.
King, J.E., 2003. Running a best-subsets logistic regression: an alternative to stepwise methods. *Educational and Psychological Measurement*, 63 (3), 392–403.
Laffan, S.W., Lubarsky, E., and Rosauer, D.F., 2010. Biodiverse, a tool for the spatial analysis of biological and related diversity. *Ecography*, 33 (4), 643–647.
Leathwick, J.R. and Whitehead, D., 2001. Soil and atmospheric water deficits and the distribution of New Zealand's indigenous tree species. *Functional Ecology*, 15 (2), 233–242.
Lin, H.S., *et al*., 1999. Effects of soil morphology on hydraulic properties: I. Quantification of soil morphology. *Soil Science Society of America Journal*, 63 (4), 948–954.
Mackey, B.G., *et al*., 1988. Assessing representativeness of places for conservation reservation and heritage listing. *Environmental Management*, 12 (4), 502–514.
Mackey, B.G., Berry, S.L., and Brown, T., 2008. Reconciling approaches to biogeographical regionalization: a systematic and generic framework examined with a case study of the Australian continent. *Journal of Biogeography*, 35 (2), 213–229.
Magness, D., Huettmann, F., and Morton, J., 2008. Using random forests to provide predicted species distribution maps as a metric for ecological inventory & monitoring programs. *In*: T. Smolinski, M. Milanova, and A.-E. Hassanien, eds. *Applications of computational intelligence in biology*. Berlin: Springer, 209–229.
Manion, G., 2009a. *NET generalised dissimilarity modeller (GDM) version 1.21*. Armidale: NSW Department of Climate Change and Water.
Manion, G., 2009b. A technique for constructing monotonic regression splines to enable non-linear transformation of GIS rasters. *In*: R.S. Anderssen, R.D. Braddock, and L.T.H. Newham, eds. *18th World IMACS congress and MODSIM09 international congress on modelling and simulation*, 13–17 July 2009, Cairns, Australia. Cairns, Australia: Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation, 2507–2513.
McKenzie, N.J., *et al*., 2000. *Estimation of soil properties using the Atlas of Australian Soils*. Canberra: CSIRO Land and Water.
McKenzie, N.J. and Hook, J., 1992. *Interpretations of the Atlas of Australian Soils*. Canberra: CSIRO Division of Soils.
Middleton, N. and Thomas, D., 1997. *World atlas of desertification*. 2nd ed. London: Edward Arnold.
Minasny, B. and Mcbratney, A.B., 2010. Methodologies for global soil mapping digital soil mapping. *In*: J.L. Boettinger, *et al*., eds. *Digital soil mapping*. Netherlands: Springer, 429–436.
Pagel, J. and Schurr, F.M., 2012. Forecasting species ranges by statistical estimation of ecological niches and spatial population dynamics. *Global Ecology and Biogeography*, 21 (2), 293–304.
Pearce, J. and Ferrier, S., 2000. An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological Modelling*, 128 (2–3), 127–147.
Pearson, R.G., *et al*., 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography*, 34 (1), 102–117.
Peterson, A.T., *et al*., 2011. *Ecological niches and geographic distributions (MPB-49)*. Princeton, NJ: Princeton University Press.
Peterson, A.T. and Nakazawa, Y., 2008. Environmental data sets matter in ecological niche modelling: an example with *Solenopsis invicta* and *Solenopsis richteri*. *Global Ecology and Biogeography*, 17 (1), 135–144.
Phillips, S.J., Anderson, R.P., and Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190 (3–4), 231–259.
Phillips, S.J. and Dudík, M., 2008. Modeling of species distributions with MaxEnt: new extensions and a comprehensive evaluation. *Ecography*, 31 (2), 161–175.
Rabus, B., *et al*., 2003. The shuttle radar topography mission: a new class of digital elevation models acquired by spaceborne radar. *ISPRS Journal of Photogrammetry and Remote Sensing*, 57 (4), 241–262.

Smith, T. and Huston, M., 1989. A theory of the spatial and temporal dynamics of plant communities. *Plant Ecology*, 83 (1), 49–69.

Thomas, K., Keeler-Wolf, T., and Franklin, J., 2002. A comparison of fine- and coarse-resolution environmental variables toward predicting vegetation distribution in the Mojave Desert. *In*: J.M. Scott, *et al*., eds. *Predicting species occurrences: issues of accuracy and scale*. Covello, CA: Island Press, 133–139.

Viscarra Rossel, R.A., 2011. Fine-resolution multiscale mapping of clay minerals in Australian soils measured with near infrared spectra. *Journal of Geophysical Research*, 116 (F4), F04023.

Western, A. and Mckenzie, N., 2004. *Soil hydrological properties of Australia version 1.0.1*. Melbourne: CRC for Catchment Hydrology.

Wilford, J., 2012. A weathering intensity index for the Australian continent using airborne gamma-ray spectrometry and digital terrain analysis. *Geoderma*, 183–184, 124–142.

Williams, K.J., 1998. *Predicting eucalypt distributions in Tasmania: an application of generalised linear modelling*. Thesis (PhD). University of Tasmania, Hobart.

Williams, K.J., *et al*., 2010. *Harnessing continent-wide biodiversity datasets for prioritising national conservation investment*. Canberra: CSIRO Ecosystem Sciences, A report prepared for the Department of Sustainability, Environment, Water, Population and Communities, Australian Government, Canberra.

Williams, K.J., Norman, P., and Mengersen, K., 2000. Predicting the natural occurrence of blackbutt and Gympie messmate in Southeast Queensland. *Australian Forestry*, 63 (4), 41–52.

Williams, K.J. and Potts, B., 1996. The natural distribution of *Eucalyptus* species in Tasmania. *Tasforests*, 8, 39–165.

Williams, K.J., *et al*. 2012. Using Bayesian mixture models that combine expert knowledge and GIS data to define ecoregions. *In*: A.H. Perera, C.A. Drew, and C.J. Johnson, eds. *Expert knowledge and its application in landscape ecology*. New York: Springer, 229–251.

Wilson, J. and Gallant, J., eds., 2000. Digital terrain analysis. *In*: *Terrain analysis: principles and applications*. New York: John Wiley, 1–27.