10

Biodiversity informatics for climate change studies

A. Culham

School of Biological Sciences and *The Walker Institute for Climate Change, University of Reading, UK*

C. Yesson

Institute of Zoology, Zoological Society of London, UK

Abstract

Modelling the impacts of climate change on biodiversity in a phylogenetic context combines the disparate disciplines of phylogenetics, geographic information systems, niche ecology and climate change research. Each subject has its own approach, literature and data. The strength of an integrative research, known as 'phyloclimatic modelling', is that it provides novel insights into the possible interactions of life and climate over millions of years. However, the risk is that problems associated with each subject area might be compounded if analyses are not conducted with care. The continuous development of analytical approaches and the steady increase in data availability have offered new opportunities for data combination. Modelling techniques and output for climate, ecological niche modelling,

Climate Change, Ecology and Systematics, ed. Trevor R. Hodkinson, Michael B. Jones, Stephen Waldren and John A. N. Parnell. Published by Cambridge University Press.

phylogeny reconstruction and temporal calibration are becoming stronger, and the reliability of results is quantifiable. In contrast, there is still a desperate lack of fundamental data on organismal distribution and on fossil history of lineages. When theories of taxonomic delimitation change, there are subsequent changes in organismal names. This creates difficulty for name-based data retrieval, but techniques are being developed to reduce this problem. Improvements in theory, associated tools and data availability will broaden the applicability of phyloclimatic modelling.

10.1 Background

Modelling the impact of climate change on the world's biota is an aspirational goal dependent on the availability of both large amounts of data and substantial computing resources. These models can be used to help us understand evolutionary relationships and ecological requirements of species, and to estimate their past, present and future distributions. The impacts of climate change on plant life are of major concern to humans because plants, apart from their intrinsic interest, play a vital role in ecosystem function and in food production and security (Heywood, 2009). The data required for modelling include species' occurrence locations, climatic variables, edaphic information and characters for phylogenetic reconstructions, while computing resources are required to build climatic and niche models and to analyse the data. The integration of these wide-ranging variables is known as 'phyloclimatic modelling' (Yesson and Culham, 2006a). There are now vast repositories of data available through distributed systems that offer the potential to allow modelling of biotic distribution patterns, phylogenies, ecological niches and the impacts of climate change without researchers having to leave their desks. However, these data should not be approached naively. Caution, and awareness of their weaknesses as well as their strengths, is needed. Before modelling can take place it is essential to consider what is being modelled. The relationships between an organism and its environment are complex, encompassing biotic and abiotic factors, functioning from microscales of a few millimetres through to macroscales of continental expanses.

One popular approach, which can be used to help understand the ecological requirements of species and estimate their distribution, is ecological niche modelling (Rödder et al., Chapter 11). Current niche modelling techniques necessarily focus on abiotic factors that show continuous variation, such as climatic conditions, and are usually referred to as models of the 'fundamental niche' (Hutchinson, 1957). Soberón (2007) reviewed niche definitions in this context, referring to these macroscale models as the 'Grinnellian niche' and explicitly excluding biotic interactions from such models. This definition is appropriate to much of the current niche modelling activity (Elith et al., 2006).

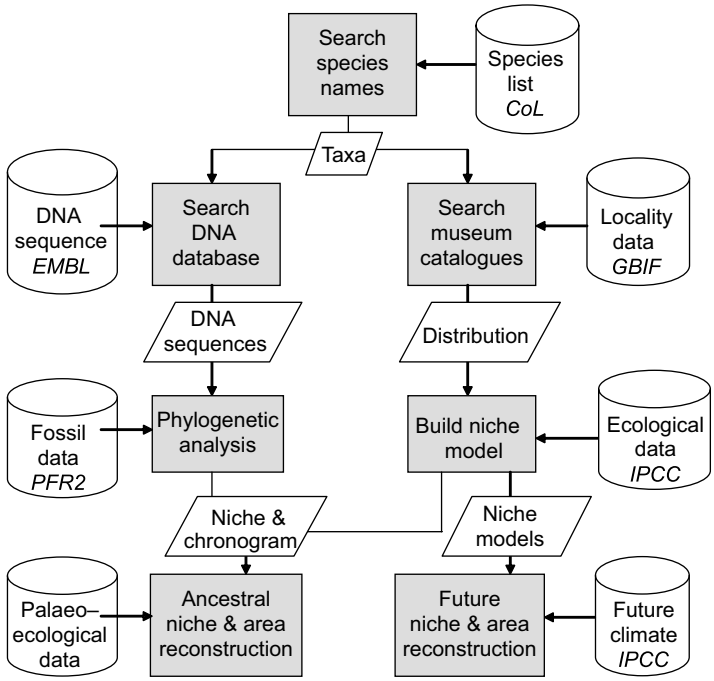

Figure 10.1 A simplified flowchart indicating data and processes for a phyloclimatic modelling workflow. Data sources are represented as cylinders, with an example database in italics. Processes are in grey boxes, outputs are in rhomboids. CoL, Catalogue of Life; GBIF, Global Biodiversity Information Facility; IPCC, Intergovernmental Panel on Climate Change; PFR2, Plant Fossil Record 2; EMBL, European Molecular Biology Laboratory.

However, in order to understand fully the interactions of species with climate, it is desirable to combine knowledge of present distribution and climate with evolutionary history (Yesson and Culham, 2006a), and hence to identify patterns for phylogenetic lineages as well as extant individual species. Such phyloclimatic modelling work requires access to substantial amounts of distributed data (Graham et al., 2004; Peterson, 2006; Yesson et al., 2007; Guralnick and Hill, 2009) and combination of these using appropriate analytical techniques (Pahwa et al., 2006). Figure 10.1 shows an example workflow for such an approach. Data on current climatic, distributional and edaphic factors are brought together to model the current bioclimatic niche of a series of species (or populations, or higher taxa). The physical data defining the niches are coded as characters on a chronogram for the same group of species (ideally based on DNA sequence data calibrated against fossil data). Reconstructions of character states at internal nodes parameterise ancestral niche models for those hypothetical taxa. The ancestral niche models are then fitted to palaeoclimate models to establish areas of potential palaeodistribution

of species. Part of the process can also be worked forward in time to estimate the possible impact of climate change on monophyla in the future. This chapter reviews some of the available data, some services that draw on those data, and quality-control issues with distributed data sources. It also highlights challenges for the future.

10.2 Biodiversity informatic data sources

10.2.1 Climate model data

Without doubt the single most focused research investment in this field is in the production of future climate models. Source climate data on which these models are based are gathered from weather stations and atmospheric probes around the world. A range of models are used (Caballero and Lynch, Chapter 2) but the predominant ones for use in predictions of global climate change are coupled atmosphere–ocean models such as HadCM3 (resolution 2.5 × 3.75 degrees latitude × longitude and 19 levels of atmosphere) and GFDL CM2.X (resolution 2.5 × 2 degrees and 25 levels of atmosphere) as adopted by the Intergovernmental Panel on Climate Change (IPCC, 2007). Such models are extremely complex and demand high-performance computing resources (Slingo et al., 2009; Washington et al., 2009) that are now at the petaflop level (a thousand trillion floating point operations per second). A new generation of massive parallel computers is allowing a bridge between climate and weather models (Slingo et al., 2009). In contrast, there is relatively little work on palaeoclimate modelling (Sellwood and Valdes, 2006; Williams et al., 2007), but this is essential if biotic evolution is to be understood in relation to climate change over evolutionary time (Yesson and Culham, 2006a, 2006b).

10.2.2 Distributional data

Perhaps the best place to look for distributional data is the Global Biodiversity Information Facility (GBIF – www.gbif.org), the largest data portal to herbarium, museum and other specimen data; in April 2009 it included *c*. 8000 data sets from *c*. 300 data providers comprising *c*. 175 000 000 occurrence records. Large figures such as this look impressive and offer a mean of nearly 100 records per named species for the roughly 1.8 million species currently recognised (Bisby et al., 2009). If we assume a minimum data requirement of between 5 and 50 independent observations (Hernandez et al., 2006; Wisz et al., 2008), and if these distributional records were spread evenly for both taxonomy and geography, then we already have the geographic data needed to attempt ecological niche modelling for almost all of the world's biota. The reality is that coverage is uneven for both taxonomic and geographic reporting (Graham et al., 2004; Yesson et al., 2007; Collen et al., 2008). Some major taxonomic groups, and many species, are completely lacking data

Table 10.1 Locality data available via the Global Biodiversity Information Facility (GBIF) in April 2009.

Kingdom	Approximate number of recorded species[a]	GBIF species[b]	GBIF occurrences[c]
Animalia	5 500 000	> 250 000	90 062 361
Archaea	n/a	320	1
Bacteria	1 000 000	11 304	34 550
Chromista	200 000	6 782	386 212
Fungi	1 500 000	103 670	1 429 696
Plantae	440 000	> 250 000	32 858 998
Protozoa	260 000	11 124	1 270 532
Viruses	400 000	277	0

[a] Species numbers from www.environment.gov.au/biodiversity/abrs/publications/other/species-numbers.

[b] GBIF species downloaded using the 'species from results' link from the kingdom occurrence summary pages. Note that this download is capped at 250 000 – this limit was reached for Animalia and Plantae.

[c] GBIF occurrences from the kingdom overview page.

(Table 10.1). For example, no georeferenced data are available for the viruses, and the entire kingdom Archaea is represented by one data point, while others, such as the class Aves (birds) represent almost half the total georeferenced data (60 261 221 records in April 2009) for only about 10 000 species (http://avibase.bsc-eoc.org/avibase.jsp), giving an impressive average in excess of 6000 records per species!

Geographic coverage shows similar patchiness. A glance at the data density maps for Plantae (plants) and Animalia (animals) suggests that Europe and North America are areas of highest biodiversity, while the rainforests of Brazil are shown as biodiversity-poor areas (Fig 10.2). It should be noted that many Brazilian data are available through the species link portal (http://splink.cria.org.br), which includes many data not currently accessible via the GBIF portal. A detailed investigation using the Fabaceae (= Leguminosae; pea and bean family) as an exemplar group shows this pattern to scale through all levels of geography and taxonomy within GBIF data (Yesson et al., 2007). Patchy coverage is combined with inconsistent data quality, something that is not surprising considering that data sources range from modern global positioning system (GPS) data accurate to a few millimetres through to museum specimens collected long before most of the world was mapped accurately and for which data have had to be interpreted and digitised manually.

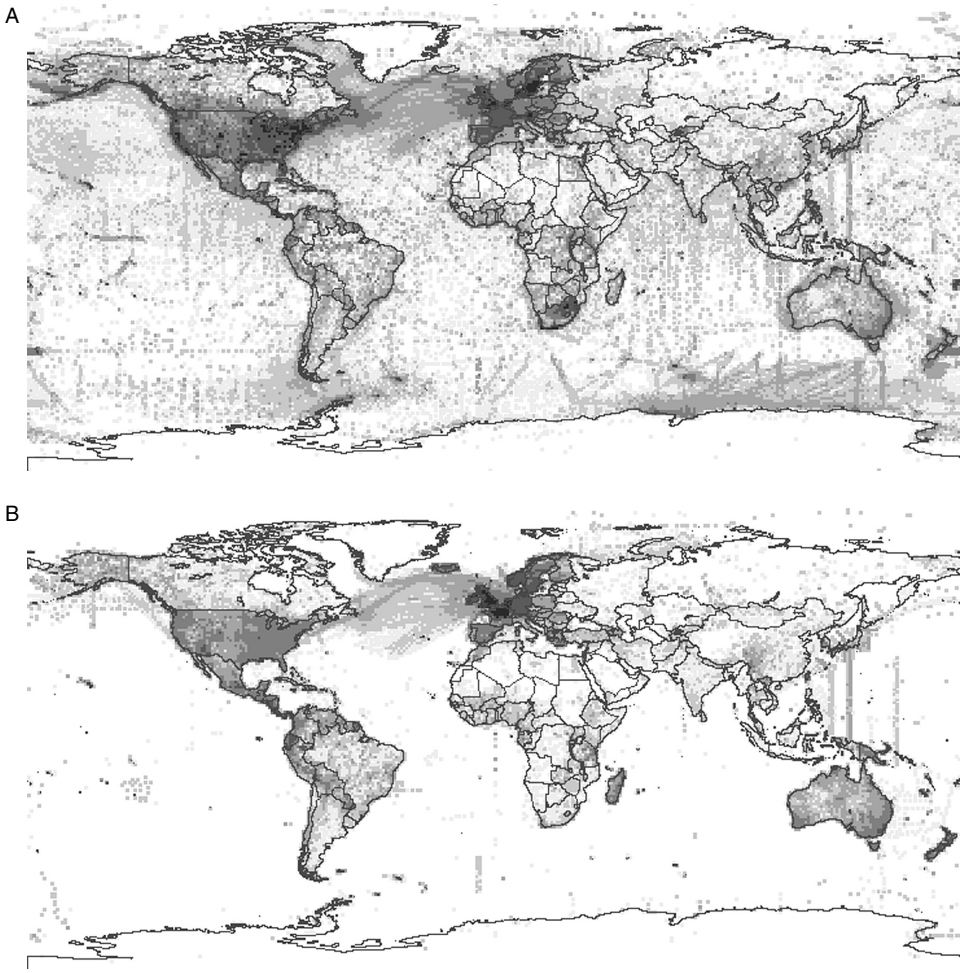

Figure 10.2 Distribution of locality data available through the Global Biodiversity Information Facility (GBIF) in April 2009 for (A) Animalia and (B) Plantae. Darker areas indicate higher frequency.

10.2.3 Taxonomic data

On top of issues with distribution data are taxonomic errors, caused by problems such as ambiguous synonyms (Page, 2005). Although organisations such as GBIF are integrating taxonomic lists into their data, there are still problems. For example, of the 21 000 data points for the tropical tree family Ebenaceae some 11 000 are in the north Atlantic Ocean, not because of problems with georeferencing, but because the family includes in its synonymy the genus *Paralia*, a name also used for a genus of phytoplankton! Blind use of such data in this case would lead to more than 53% of the distribution data being attributed wrongly. Other cases may be less obvious. However, new developments such as the Life Science

Identifier (LSID) (Clark et al., 2004) and taxonomically intelligent network services (Patterson et al., 2006) may help to reduce these difficulties.

10.2.4 DNA sequence data

Phylogenetic studies are allowing patterns of change in lineages, rather than just species, to be investigated. Such research is based on phylogenetic trees predominantly built using DNA sequence data. Not only is DNA sequencing commonly used to establish the relationships among morphological species, it is also now widely used to assess species- and populational-level boundaries that might be invisible using purely morphological data (e.g. Hartmann et al., 2006; Leliaert et al., 2009; Bateman, Chapter 3; Bernardo, Chapter 18). DNA sequence data are accessed via the three data portals for molecular data: Genbank (www.ncbi.nlm.nih.gov/genbank), EMBL-BANK (www.ebi.ac.uk/embl) and the DNA Data Bank of Japan (DDBJ) (www.ddbj.nig.ac.jp). These data portals share data, so each of the three underlying databases is largely similar in content. These databases were established during the early days of DNA sequencing and were in place for subsequent large-scale sequencing work. Many scientific journals adopted a requirement for authors to deposit data in these databases before publication of their results, and that obligation has ensured a comprehensive record of DNA sequencing activity. Quality control of data is primarily dependent on the data provider, although submissions are reviewed by specialist teams for the receiving database to ensure appropriate attempts at annotation of sequence data. There remains the problem that explanation of the sequencing method, data reliability and species/sample authentication is reported in the source publication rather than recorded in the database. This century, at least, the opportunity to cite voucher specimens that allow independent authentication of identification has become more common. Data quality, assessed through DNA sequencing electronic trace files, is an ongoing issue because of the large storage size of trace files versus text-based DNA sequence files. Trace files show the quality of the DNA read and not just the sequence of bases in the recorded DNA sequence. Trace files are now being stored for several specialist projects, such as whole genome studies and expressed sequence tag (EST) studies (e.g. www.ncbi.nlm.nih.gov/Traces and http://trace.ensembl.org), but the number of records is still trivial in comparison with the number of text-based sequence depositions.

10.2.5 Fossil data

The least complete source of data for phyloclimatic modelling studies, by far, is that for the fossil record. There are two main online databases for macrofossil data: the Paleobiology Database (http://paleodb.org), which offers a form-based search, and the Fossil Record (www.fossilrecord.net), which offers a series of downloadable files organised by taxonomic groups. In addition, there are other specialist databases such as the Fossil Pollen Database (http://pollen.cerege.fr/fpd-epd), the

Palaeoflora database (www.palaeoflora.de), which specifically includes climate preference data for the closest living relatives of fossil species, and Chronos (www.chronos.org), a data portal for data sets covering geological timescales. However, data are still scattered over a range of websites and in a broad variety of formats.

10.3 Tools

10.3.1 Niche modelling

There is a range of competing algorithms for ecological niche modelling. One of the earliest developed and most straightforward is BIOCLIM (Busby, 1991), but others have developed the approach using genetic algorithms (Stockwell and Peters, 1999), maximum entropy (Phillips et al., 2006) and many others (Elith et al., 2006). Some of these approaches have dedicated software such as DesktopGarp (www.nhm.ku.edu/desktopgarp) and Maxent (www.cs.princeton.edu/~schapire/maxent), but there are also modelling packages that offer a range of algorithms such as OpenModeller (http://openmodeller.sourceforge.net) within one desktop interface. A strength of these packages is that they are offered free of charge for research use, and in the case of OpenModeller as an open source project with a team of contributors around the world.

10.3.2 Online systems

Several online facilities are available that give an idea of the potential for future web-based biodiversity services. There have been two contrasting approaches to providing the computing power for niche modelling using distributed data. One is the provision of an application on a dedicated server via a web interface: for example, the model used for WhyWhere (http://landshape.org/enm/whywhere-20-server). The other is the use of distributed computing via a web interface and through a managing server that distributes jobs to desktop PCs: for example, the system for Lifemapper (www.lifemapper.org). Both systems allow use of GBIF distribution data, but both are limited in their niche modelling approaches when compared with desktop software such as OpenModeller. These systems begin to show the opportunities given by large distributed data sets. However, they continue to be reliant on trust in the quality and consistency of those data and still require substantial human input for large modelling projects.

10.4 Conclusions: present uses and future needs

Phyloclimatic modelling approaches have already been used to investigate climate-related evolution and distribution in several genera. In the Mediterranean basin, speciation in *Anthemis* has been linked to aridification 9 million years ago (mya)

and to climatic oscillation in the past 3.5 million years (Lo Presti and Oberprieler, 2009), while *Cyclamen* appears to have been influenced more by geographic separation caused by fluctuating sea level over a similar period (Yesson et al., 2009). The Pacific Northwest mesic forest organisms of North America have been studied through palaeo-niche modelling to better understand current biogeography in the light of putative palaeogeographical distributions during cycles of glaciation (Carstens and Richards, 2007). Niche evolution over phylogenetic time has been applied to a range of terrestrial species and genera including Icteridae (American blackbirds – Eaton et al., 2008), *Oenothera* (evening primroses – Evans et al., 2009), *Drosera* (sundews – Yesson and Culham, 2006a) and Poaceae (grasses – Jakob et al., 2009) as well as marine algae (Verbruggen et al., 2009). These papers highlight the potential of a phyloclimatic approach to gain insights into, and deeper understanding of, biogeography and evolutionary history. They provide examples of the high explanatory power of past distributions on present ones over long (Yesson and Culham, 2006a) and intermediate (Carstens and Richards, 2007) timescales. Yesson and Culham (Chapter 12) outline how phyloclimatic modelling approaches have been applied to study genera in the Mediterranean-type climatic zones of Australia and the Mediterranean basin. Knowledge of the past informs plans for the future.

For the future, several developments are under way. For example, integrated data pipelines that allow experimental modelling systems to be automated over large numbers of species are in development. The Kepler project (https://kepler-project.org/users/projects-using-kepler) is an example using workflow management tools. The success of developments in such integrative science bringing together taxonomy, ecology, climatology and computer science will ultimately rest on the security of research funding in this area and on the development of open source tools that can be built collaboratively on an international scale.

Acknowledgements

We wish to thank the Biotechnology and Biological Sciences Research Council for funding the BioDiversity World project, the University of Reading for funding the second author's PhD, and numerous colleagues for feedback and discussion of our ideas over the past five years.

References

Bisby, F., Roskov, Y., Orrell, T. et al. (2009). Species 2000 and ITIS Catalogue of Life: 2009 Annual Checklist. Reading: Species 2000. www.Catalogueoflife.Org/annual-checklist/2009.

Busby, J. R. (1991). BIOCLIM: a bioclimatic analysis and prediction system. In

Nature Conservation: Cost Effective Biological Surveys and Data Analysis, ed. C. R. Margules and M. P. Austin. Melbourne: CSIRO, pp. 64–68.

Carstens, B. C. and Richards, C. L. (2007). Integrating coalescent and ecological niche modeling in comparative phylogeography. *Evolution*, **61**, 1439–1454.

Clark, T., Martin, S. and Liefeld, T. (2004). Globally distributed object identification for biological knowledgebases. *Briefings in Bioinformatics*, **5**, 59–70.

Collen, B., Ram, M., Zamin, T. and McRae, L. (2008). The tropical biodiversity data gap: addressing disparity in global monitoring. *Tropical Conservation Science*, **1**, 75–88.

Eaton, M. D., Soberon, J. and Peterson, T. (2008). Phylogenetic perspective on ecological niche evolution in American blackbirds (family Icteridae). *Biological Journal of the Linnean Society*, **94**, 869–878.

Elith, J., Graham, C. H., Anderson, R. P. et al. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Evans, M. E. K., Smith, S. A., Flynn, R. S. and Donoghue, M. J. (2009). Climate, niche evolution, and diversification of the 'Bird-cage' Evening primroses (*Oenothera*, sections *Anogra* and *Kleinia*). *American Naturalist*, **173**, 225–240.

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C. and Peterson, A. T. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution*, **19**, 497–503.

Guralnick, R. and Hill, A. (2009). Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. *Bioinformatics*, **25**, 421–428.

Hartmann, F. A., Wilson, R., Gradstein, S. R., Schneider, H. and Heinrichs, J. (2006). Testing hypotheses on species delimitations and disjunctions in the liverwort *Bryopteris* (Jungermanniopsida: Lejeuneaceae). *International Journal of Plant Sciences*, **167**, 1205–1214

Hernandez, P. A., Graham, C. H., Master, L. L. and Albert, D. L. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, **29**, 773–785.

Heywood, V. H. (2009). The impacts of climate change on plant species in Europe, with contributions by A. Culham. Strasbourg: Convention on the Conservation of European Wildlife and Natural Habitats Standing Committee.

Hutchinson, G. E. (1957). Concluding remarks. Cold Spring Harbor Symposium. *Quantitative Biology*, **22**, 415–427.

Intergovernmental Panel on Climate Change (IPCC) (2007). *Climate Change 2007: Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, ed. R. K. Pachauri and A. Reisinger. Geneva: IPCC.

Jakob, S. S., Martinez-Meyer, E. and Blattner, F. R. (2009). Phylogeographic analyses and paleodistribution modeling indicate Pleistocene in situ survival of *Hordeum* species (Poaceae) in southern Patagonia without genetic or spatial restriction. *Molecular Biology and Evolution*, **26**, 907–923.

Leliaert, F., Verbruggen, H., Wysor, B. and De Clerck, O. (2009). DNA taxonomy in morphologically plastic taxa: algorithmic species delimitation in the *Boodlea* complex (Chlorophyta: Cladophorales). *Molecular Phylogenetics and Evolution*, **53**, 122–133.

Lo Presti, R. M. and Oberprieler, C. (2009). Evolutionary history, biogeography and eco-climatological differentiation of the genus *Anthemis* L. (Compositae, Anthemideae) in the circum-Mediterranean area. *Journal of Biogeography*, **36**, 1313–1332.

Page, R. D. M. (2005). A taxonomic search engine: federating taxonomic databases using web services. *BMC Bioinformatics*, **6**, 48.

Pahwa, J. S., Jones, A. C., White, R. J. et al. (2006). Supporting the construction of workflows for biodiversity problem-solving accessing secure, distributed resources. *Scientific Programming*, **14**, 195–208.

Patterson, D. J., Remsen, D., Marino, W. A. and Norton, C. (2006). Taxonomic indexing: extending the role of taxonomy. *Systematic Biology*, **55**, 367–373.

Peterson, A. T. (2006). Uses and requirements of ecological niche models and related distributional models. *Biodiversity Informatics*, **3**, 59–72.

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

Sellwood, B. W. and Valdes, P. J. (2006). Mesozoic climates: general circulation models and the rock record. *Sedimentary Geology*, **190**, 269–287.

Slingo, J., Bates, K., Nikiforakis, N. et al. (2009). Developing the next-generation climate system models: challenges and achievements. *Philosophical Transactions of the Royal Society of London A*, **367**, 815–831.

Soberón, J. (2007). Grinnellian and Eltonian niches and geographic distributions of species. *Ecological Letters*, **10**, 1115–1123.

Stockwell, D. R. B. and Peters, D. P. (1999). The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographic Information Systems*, **13**, 143–158.

Verbruggen, H., Tyberghein, L., Pauly, K. et al. (2009). Macroecology meets macroevolution: evolutionary niche dynamics in the seaweed *Halimeda*. *Global Ecology and Biogeography*, **18**, 393–405.

Washington, W. M., Buja, L. and Craig, A. (2009). The computational future for climate and earth system models: on the path to petaflop and beyond. *Philosophical Transactions of the Royal Society of London A*, **367**, 833–846.

Williams, M., Haywood, A. M., Gregory, J. and Schmidt, D. N. (2007). *Deep-time Perspectives on Climate Change: Marrying the Signal From Computer Models and Biological Proxies*. London: Geological Society of London on behalf of the Micropalaeontological Society.

Wisz, M. S., Hijmans, R. J., Li, J. et al. (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, **14**, 763–773.

Yesson, C. and Culham, A. (2006a). Phyloclimatic modelling: combining phylogenetics and bioclimatic modelling. *Systematic Biology*, **55**, 785–802.

Yesson, C. and Culham, A. (2006b). A phyloclimatic study of *Cyclamen*. *BMC Evolutionary Biology*, **6**, 72.

Yesson, C., Brewer, P. W., Sutton, T. et al. (2007). How global is the Global Biodiversity Information Facility? *PLoS ONE,* **2**, e1124.

Yesson, C., Toomey, N. H. and Culham, A. (2009). *Cyclamen*: time, sea and speciation biogeography using a temporally calibrated phylogeny. *Journal of Biogeography*, **36**, 1234–1252.