

Final report: CEBRA 1402B Tools and approaches for invasive species distribution modelling for surveillance

Simon Barry^A, Jane Elith^B, Daniel Heersink^A, Peter Caley^A, Mike Kearney^B, Phil Tenant^C, Tony Arthur^C

^ACSIRO Health & Biosecurity

^BUniversity of Melbourne

^CABARES



Contents

CONTENTS	3
FIGURES	4
TABLES	7
1 EXECUTIVE SUMMARY	8
2 INTRODUCTION	9
2.1 PROJECT BACKGROUND	9
2.2 THE ISSUE	9
2.3 PROJECT OUTLINE	11
3 REVIEW OF PROXIMAL VARIABLES	13
3.1 PREAMBLE	13
3.2 DO WE KNOW WHICH PREDICTORS ARE PROXIMAL, BASED ON THEORY?	13
3.3 SPATIAL SCALE	14
3.4 SOURCES OF GLOBAL TERRESTRIAL DATASETS FOR PREDICTOR VARIABLES	15
3.5 HOW HAS THEORY BEEN TRANSLATED INTO PRACTICE, IN CHOICE OF PREDICTOR VARIABLES?	16
3.7 DOES IT MATTER WHICH PREDICTORS ARE CHOSEN?	18
3.8 DEALING WITH RELATIONSHIPS BETWEEN PREDICTOR VARIABLES	19
3.9 DISCUSSION	20
4 OUTLINE OF MODELLING PROCESS	21
5 EMPIRICAL IDENTIFICATION OF PROXIMAL VARIABLES	25
5.1 PROTOCOL DEVELOPMENT FOR MODEL FITTING	25
5.2 CASE STUDIES	28
5.3 PROTOCOL DEVELOPMENT FOR MODEL FITTING – CASE STUDIES	28
5.3.1 Fire ant (<i>Solenopsis invicta</i>)	29
5.3.2 Asian gypsy moth (<i>Lymantria dispar</i>)	36
5.3.3 The invasive / oriental fruit fly (<i>Bactrocera invadens</i> , <i>B. dorsalis</i> , <i>B. papayae</i> , and <i>B. philippinensis</i>)	43
5.3.4 Myrtle/guava rust (<i>Puccinia psidii</i> s.l.)	50
5.3.5 Cane toads (<i>Rhinella marina</i>)	55
6 PROBLEMS WITH PROJECTIONS	61
7 SYNTHESIS	73
7.1 DISCUSSION	73
7.2 PROTOCOL	75
7.3 FUTURE DEVELOPMENTS	76
8 REFERENCES	77
9 APPENDICES	I
9.1 APPENDIX A1: LITERATURE REVIEW OF PAPERS RELATED TO PREDICTOR VARIABLES AND SPECIES DISTRIBUTIONS	I
9.2 APPENDIX A2: THE 35 BIOCLIMATIC VARIABLES IN ANUCLIM	I

Figures

Figure 1. Distribution of fire ants (<i>Solenopsis invicta</i>) in South America (native range) and the United States (introduced). Data sourced from Fitzpatrick <i>et al.</i> (2007).	10
Figure 2. Example of projection problems. Left hand panel is the mean habitat suitability for fire ants (<i>Solenopsis invicta</i>) based on their native South American distribution (average of 10 folds) utilising all BIOCLIM variables. Right hand panel is standard deviation of the 10 folds. Data sourced from Fitzpatrick <i>et al.</i> (2007).	11
Figure 3. Known locations for <i>Halictus smaragdulus</i> (black & grey dots), and Maxent predictions in native range for models fitted to (a) 19 (c) 4 and (e) 2 climate variables (see text).	18
Figure 5 Effect of dimension on the distance to the nearest neighbour for a fixed sample size.	23
Figure 6. Relationship between July maximum temperature and the ecoclimatic temperature index	26
Figure 7. Best fitting GAM using BIOCLIM temperature (BIO3) and precipitation (BIO18) for fire ants (<i>Solenopsis invicta</i>) with continental scale background. Blue crosses in South America denote occurrences. Aqua dots in the USA denote the invaded range (not used in model development).	32
Figure 8. Projection of best fitting GAM using BIOCLIM temperature (BIO3) and precipitation (BIO18) for red imported fire ant to Australia and New Zealand. Background is at continental scale.	33
Figure 9. Best fitting GAM using MICROCLIM temperature (O3.0) and precipitation (BIO18) for fire ants (<i>Solenopsis invicta</i>) with continental scale background. Blue crosses in South America denote occurrences. Blue dots in the USA denote the invaded range (not used in model development).	33
Figure 10. Projection of best fitting GAM using BIOCLIM temperature (BIO3) and precipitation (BIO18) for red imported fire ant to Australia and New Zealand. Background is at local scale.	34
Figure 11. Projection of best fitting GAM using expert chosen variables for red imported fire ant to Australia and New Zealand. Background is continental.	35
Figure 12. Best fitting GAM to BIOCLIM temperature (BIO1) and precipitation (BIO12) for Asian gypsy moth (<i>Lymantra dispar</i>). Blue crosses are recorded locations used in model development with a continental scale background.	39
Figure 13. Projection of best fitting GAM using BIOCLIM temperature (BIO1) and precipitation (BIO12) for Asian Gypsy Moth to Australia and New Zealand. Continental background.	40
Figure 14. Project alpha-hull model using of best fitting GAM parameters BIOCLIM 01 (Annual Mean Temperature) and BIOCLIM 12 (Annual Precipitation) for Asian Gypsy Moth. Continental background.	41
Figure 15. Projection of expert based GAM using BIO4 (Temperature Seasonality), BIO5 (Maximum Temperature of Warmest Week), BIO6 (Minimum Temperature of Warmest Week), BIO12 (Annual Precipitation), and BIO28 (Annual Mean Moisture Index) for Asian Gypsy Moth to Australia and New Zealand. Continental background.	42
Figure 16. Best GAM derived from first 19 BIOCLIM variables. Continental background.	42
Figure 17. Distribution records for <i>B. invadens</i> used by De Meyer <i>et al.</i> (2010).	46
Figure 18. From Stephens <i>et al.</i> (2007), the distribution records for <i>Bactrocera dorsalis</i> used in their modelling (circles, native range; crosses; invaded) – as then considered a separate species to <i>B. invadens</i>	46

Figure 19. Best fitting GAM using BIOCLIM O6 (Min. Temp. of Coldest Period) and BIOCLIM 12 (Annual Precipitation) for the global distribution of the <i>Bactrocera dorsalis</i> complex. Blue crosses are reported collections. Background is continental.	47
Figure 20. Projection of best fitting GAM using BIOCLIM O6 (Min. Temp. of Coldest Period) and BIOCLIM 12 (Annual Precipitation) for the <i>Bactrocera dorsalis</i> species complex to Australia and New Zealand. Continental background.	48
Figure 21. Alpha hull model based on best fitting GAM variables BIOCLIM O6 (Min. Temp. of Coldest Period) and BIOCLIM 12 (Annual Precipitation).	49
Figure 22. Bounding box model based on best fitting GAM variables BIOCLIM O6 (Min. Temp. of Coldest Period) and BIOCLIM 12 (Annual Precipitation).	49
Figure 23. Best fitting GAM for myrtle rust based on BIOCLIM variables BIO11 and BIO14.	53
Figure 24. Best fitting GAM for myrtle rust based on BIOCLIM variables BIO11 (Mean temp. of coldest quarter) and BIO14 (Precipitation of driest period) applied to Australia and New Zealand. Observations are small blue crosses.	53
Figure 25. Alpha-hull myrtle rust model based on BIOCLIM variables BIO11 (Mean temp. of coldest quarter) and BIO14 (Precipitation of driest period). Observations are small blue crosses.	54
Figure 26. Projected suitability for myrtle rust from GAM based expert identified variables.	54
Figure 27. Best fitting GAM for the cane toad (<i>Rhinella marina</i>) based on BIOCLIM variables BIO6 (Min. Temp. of Coldest Period) and BIO18 (Precipitation of Warmest Quarter). Observations are blue crosses. Background is continental.	57
Figure 28. Projection of best fitting GAM for the cane toad (<i>Rhinella marina</i>) based on BIOCLIM variables BIO6 and BIO18 and continental background. Model is fitted to native range only.	58
Figure 29. Alpha hull projection for cane toad, BIO06 and BIO18 (best fit GAM), continental background. Note the area of predicted suitability in South America that is inhabited by the congener <i>Rhinella schneideri</i>	59
Figure 30. Alpha hull projection for cane toad, BIO06 and BIO18 (best fit GAM), continental background. .	59
Figure 31. Bounding box projection for the cane toad, BIO06 and BIO18, continental background.	60
Figure 32. Bounding box projection for the cane toad, expert derived BIOCLIM variables, continental background.	60
Figure 33. Simulated distribution, BIOCLIM variables 2 and 19.	63
Figure 34 Actual niche. The block of + symbols show the presences, and small dots are absences. Note the axes are not scaled identically in the two panels.	63
Figure 35 Empirical niche – observations shown along the selected variables, 9 and 14. Legend as above..	64
Figure 36. Comparison of predictions from model fitted to South American data to model fitted to Australian data, to all environments on the Australian continent.	64
Figure 37 Predictions and actual data for South American data	65
Figure 38. Predictions from native range model, Australian model and actual data.	65
Figure 39. ROC curves from native range model and for projections to Australia.	66
Empirical niche – BIOCLIM 5 and BIOCLIM 16Figure 40. Simulated distribution, BIOCLIM variables 2 and 19.	67
Figure 41. Actual niche.	67
Figure 42. Empirical niche.	68

Figure 43. Comparison of predictions from model fitted to South American data to model fitted to Australian data. Predictions based on Australian environmental data.	69
Figure 44. Predictions and actual data for South American data.	69
Figure 45. Predictions from native range model, Australian model and actual data.	70
Figure 46. ROC curves from native range model and for projections to Australia.	70
Figure 47. 100 randomly sampled models, ROC curve based on fit and projection to Australia.	71
Figure 48 100 randomly sampled models, ROC curve based on comparison of projection to Australia against truth observations in Australia.	72

Tables

Table 1. Examples of proposed variables driving distributions of various groups of organisms.	14
Table 2. Global GIS data for predictor variables in models.....	17
Table 3. BIOCLIM bioclimatic variables and their description.	27
Table 4. A selection of species distribution models from the literature for fire ants (<i>Solenopsis invicta</i>).....	29
Table 5. Gypsy moth (<i>Lymantria dispar</i>) models in the literature.	37
Table 6. Oriental fruit fly (<i>Bactrocera dorsalis</i> complex) models in the literature.	44
Table 7. Myrtle rust (<i>Puccinia psidii sensu lato</i>) models in the literature.....	50
Table 8. Cane toads (<i>Rhinella marina</i>) models in the literature.	55

1 Executive summary

Risk-based biosecurity surveillance systems rely, amongst other things, on reliable spatial models of the potential habitat of species of concern, often termed species distribution models (SDMs). A SDM can be used for a variety of purposes:

- By understanding the possible extent of an incursion it can underpin the estimation of the potential economic, environmental and human health costs.
- By combining an SDM with pathway analysis it can assist in prioritizing where surveillance effort should be expended to maximise the likelihood of early detection.

An understanding of the potential habitats the species can establish in can support eradication campaigns. A challenge in developing SDM's is the diversity of techniques and opinion in the scientific literature. No one modelling technique has emerged as being suitable for all applications. In addition, while the importance of basing models on good predictive variables is understood conceptually there has been little attempt to review this or develop concrete protocols to identify these variables.

This project reviews the available environmental data and explores the information in the literature defining proximal variables, which are variables that are most directly and closely linked to the biological process and therefore likely to be the best predictors of potential distribution. The review identified that there is not a consistent approach within ecology to identifying proximal variables. While proximal variables are well defined conceptually, identifying them from observations of the species is more complex, because correlation can obscure causation. In other words, it is difficult to measure promixity. Thus while a variable may appear strongly predictive, its performance in other locations cannot be unambiguously predicted.

Based on this review the project then explored several approaches to developing predictive models. We developed methods to try to identify proximal variables statistically using small two variable models to guard against over-fitting. We also tested new microclimatic variables with a more physiological basis than those commonly used. Once variables were selected, we then tested a range of methods for making predictions to new regions, including (1) from a fitted model, (2) from climate envelopes constructed on the selected variables. The reason for the latter is that there are a number of compelling reasons that probability based predictions to new regions will be unreliable. These methods where then applied to five case studies using pests that have or could establish in Australia and/or New Zealand. There was no clear preferred approach from this analysis.

The basis of this result was explored via simulation analysis. This analysis demonstrated that variables could be strongly predictive in the native range but weakly predictive when projected to new locations. This demonstrated a fundamental limitation in our ability to accurately perform these projections with any of the tested models.

Based on these analyses a protocol was developed that reflects this inherent uncertainty. This protocol recommends experts based assessment of proximal variables and incorporation of this uncertainty into the analysis. It recommends the use of envelopes rather than probability methods when projecting to new locations.

2 Introduction

2.1 Project background

The design of risk-based biosecurity surveillance systems rely, amongst other things, on statistically reliable spatial models of the potential habitat of species of concern, often termed species distribution models (SDMs). This reliance is because risk based calculations depend critically on being able to reasonably assess the likelihood of events.

An SDM can be used for a variety of purposes:

- By understanding the possible extent of an incursion it can underpin the estimation of the potential economic, environmental and human health costs.
- By combining an SDM with pathway analysis (Heersink *et al.* 2015) it can assist in prioritizing where surveillance effort should be expended to maximise the likelihood of early detection.
- By understanding the potential habitats the species can establish in, it can support eradication campaigns.

This project explores developing a structured approach to developing SDMs to support the mapping of the potential distribution of new pests and diseases that may be used to facilitate better biosecurity decision making. It aims to synthesise best practice approaches from available techniques in the scientific literature to provide an objective protocol that can be confidently applied and justified in decision making processes.

A range of tools for habitat suitability modelling have been developed in the academic literature (e.g. Kriticos *et al.* 2005) and some have been adapted for general deployment by biosecurity agencies. Despite these developments, CEBRA Project 1302 established that there is no single, best approach to predicting invasive species distributions. Correlation based methods are not ideally suited to predicting the distributions of pests in new environments (Elith *et al.* 2010) due to data and knowledge limitations. Other, physiologically and ecologically based approaches (e.g. Kearney and Porter 2009) may require data or understanding that are typically not available for many species. There is a gap in the form of guidelines and concrete protocols for dealing with the full set of contingencies that face biosecurity managers. This project explores developing a set of protocols to make robust and therefore defensible predictions about the expected distribution of new species in Australia and New Zealand.

2.2 The issue

The difficulties in characterising and projecting distributions of species between native and introduced ranges are reasonably well known in the scientific literature (Elith and Leathwick 2009). What is less clear is the reasons that they occur and how they can be ameliorated.

Species distribution models fitted to species data collected from the native range using available bioclimatic variables typically don't reliably project well to new environments, despite the use of standard model selection approaches aimed at avoiding model over-fitting. For example, consider the fire ant (*Solenopsis invicta*), native to South America, well established in the USA, and recently established in Australia. In Figure 2, the result of fitting a generalised additive model (GAM) to the distribution of fire ants in their native range in South America (Figure 1) using the standard 19 BIOCLIM variables and projecting to the rest of the world is illustrated. The suitability shown is the mean of the best models selected from 10 folds of the training data ($n=74$ from Fitzpatrick *et al.* (2007)). Each selected model is based on penalised regression splines, with the degree of each term restricted to no more than two (only monotonic or concave/convex relationships permitted to reduce overfitting). Despite these efforts to avoid over-fitting within the training data, the projections do not accord well with what we know about fire ant distribution (e.g. the extent of

invasion in the southern states of the USA, or their demographic vigour in Brisbane), or what we think we know (e.g. that Siberia is probably not suitable).

Some (e.g. Sutherst and Maywald 2005) have argued that fire ants are not climate-limited in their native South America, though go on to fit climate-based models to their invaded range in the southern United States. We would argue that for many species of interest, such invaded ranges are not available and the process based arguments are often untestable. Furthermore, whether it would have been apparent to researchers that fire ants were not climate-limited in the native range in the absence of knowledge of the invaded range will never be known (the counterfactual). Incorporation of invaded range in modelling is recommended in some situations (Broennimann and Guisan 2008), and there are some good conceptual reasons to include records from invaded ranges where the species has been resident for enough time to disperse throughout the landscape to suitable environments, and to persist (Elith in press). However we would argue that it would be preferable to find robust methods that are able to perform on the native range alone, as this is the data that biosecurity authorities will be most likely dealing with. More broadly, our fire ant example highlights a number of the issues, in particular the failure of these models to reliably extrapolate to new environments.

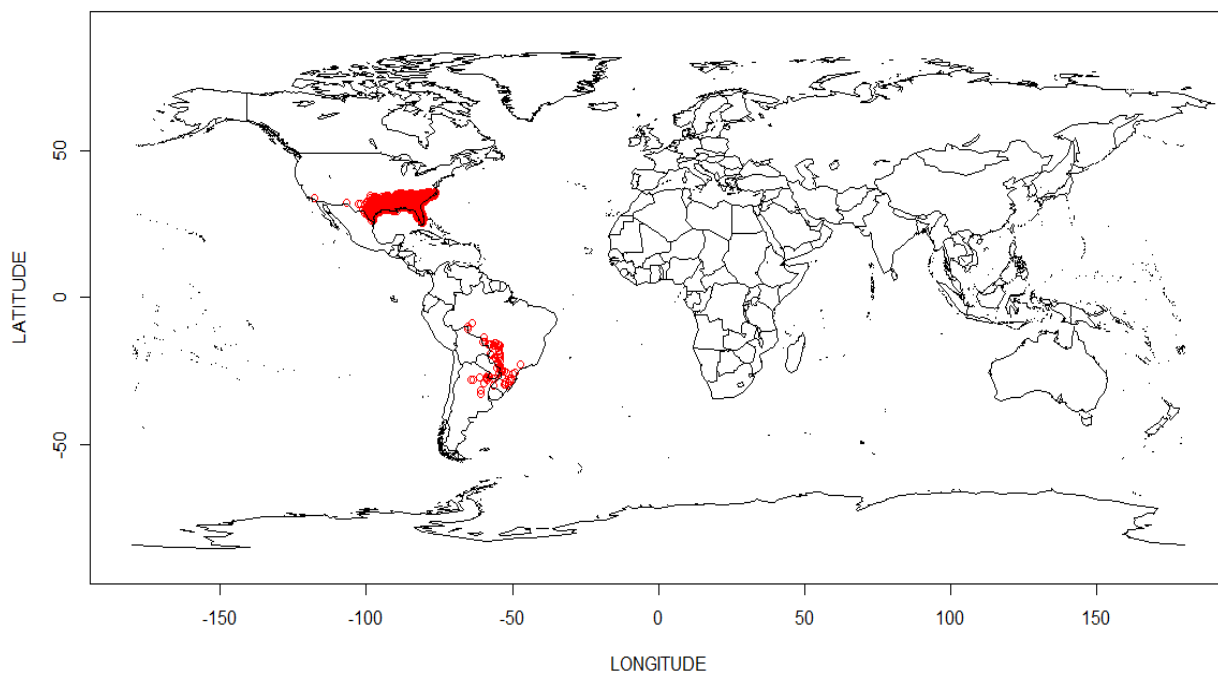


Figure 1. Distribution of fire ants (*Solenopsis invicta*) in South America (native range) and the United States (introduced). Data sourced from Fitzpatrick *et al.* (2007).

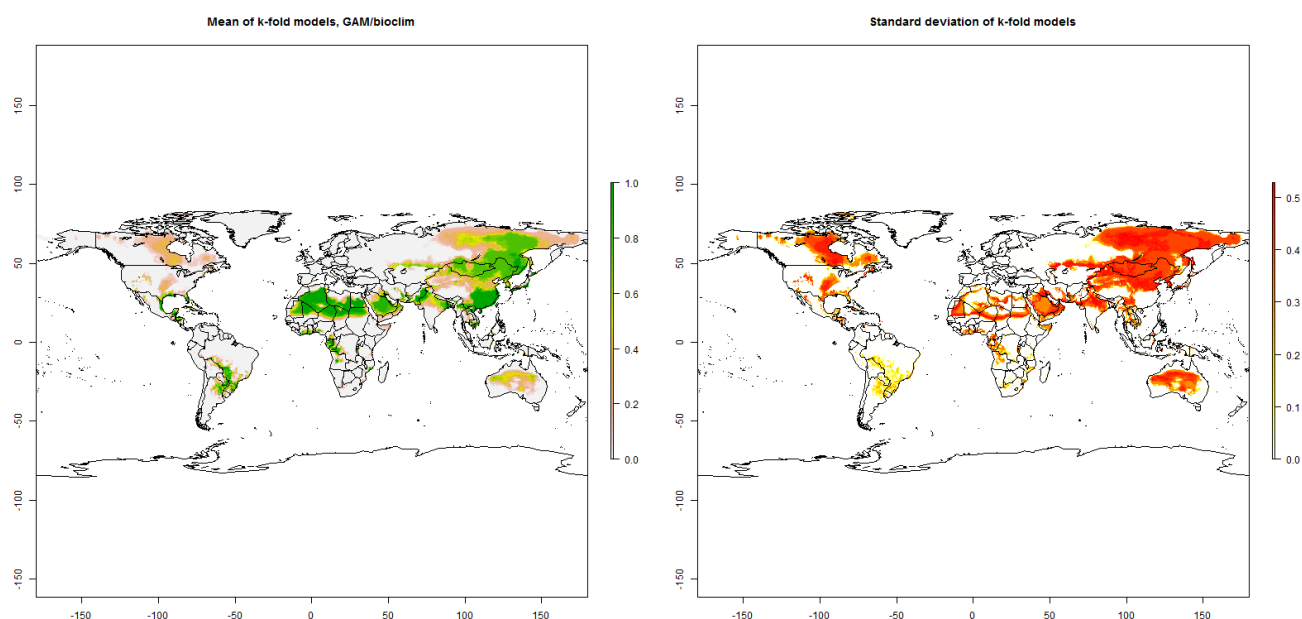


Figure 2. Example of projection problems. Left hand panel is the mean habitat suitability for fire ants (*Solenopsis invicta*) based on their native South American distribution (average of 10 folds) utilising all BIOCLIM variables. Right hand panel is standard deviation of the 10 folds. Data sourced from Fitzpatrick *et al.* (2007).

2.3 Project outline

This report explores approaches to habitat suitability modelling for non-native species with the goal of developing structured guidelines and protocols to assist managers to identify the most appropriate tools and approaches for specific applications.

This project arose from the results of a previous CEBRA project. CEBRA Project 1302 identified a number of key issues in predicting species distributions were identified:

- That simple correlative approaches would often fail because the pattern of correlation between environmental variables and distribution was not, in general fixed. A key challenge was therefore to identify proximal variables. Proximal variables (Austin 2002) are variables that are “close” to the processes that determine a species’ ecological functions and are therefore more likely to be closely associated with its distribution. Therefore a key component of this project was to gain knowledge about and approaches to identifying proximal variables.
- That probability based predictions were theoretically difficult to justify because the underlying statistical populations (native vs invaded ranges) were not equivalent. Thus approaches using techniques such as climate envelopes could be useful, because these simplify predictions into questions about whether climates in the target regions are suitable or not.

These issues led to the following work program:

1. Develop protocols to determine key predictive variables that determine an organism’s range.

This will involve:

- Reviewing literature and identifying approaches to determining predictors that will be applicable generally. These will typically be primary environmental variables.

- Reviewing modelling approaches to determine if and how they can be used to identify key environmental predictors.
- Assessing the role of expert opinion in identifying key variables and developing protocols for doing this.

2. Develop protocol/methods to project the information determined from the native range into the new domain.

This will involve the following tasks:

- Reviewing methods for constructing environmental matches from variables determined in step 1 and information on native range.
- Developing protocols and implement techniques in R, if appropriate, to develop suitable products.
- Assessing the role of expert opinion in finalising distribution and developing protocols for doing this.

As the project proceeded a number of challenges arose. In particular the discussion and identification of proximal variables in the literature was less developed than anticipated. This was communicated to the steering committee and lead to more focus on automated approaches to identifying proximal variables than was originally intended. We present the results as follows. Chapter 3 reviews the literature on proximal variables. Chapter 4 considers the impact of the curse of dimensionality on the construction of climate envelopes. Chapter 5 develops empirical approaches to identifying proximal variables. Chapter 6 explores the theoretical basis of the results. Chapter 7 discusses the results and proposes a protocol.

3 Review of proximal variables

3.1 Preamble

This section addresses the problem of choosing predictor variables for use in models predicting the potential distributions of invasive species in a new region. For this project that new region is Australia and/or New Zealand (ANZ). Prediction of potential distributions of species is a specific use of species distribution models (SDMs) that extends beyond their initial intended use (which was understanding or predicting the distribution of a species within the region in which it has been observed). Because it is an extension, this chapter will particularly target literature and ideas that have been tested in the specific situations that are relevant – i.e., where models were used to predict to new geographic localities. The emphasis on predicting to new places means that the predictors we are most interested in are the **environmental** ones. Geographic predictors can also be used in distribution modelling, but they are most relevant when we are interested in spatial contagion or other geographic processes leading to spatial autocorrelation in species occurrences. This might be useful for modelling the spread of invasive species in a new range, but not for predicting potential distributions in new regions.

Potential distributions can only be correctly predicted if the fitted model (based on records in the native range or native plus long-invaded ranges) is also relevant to the new region (e.g., ANZ). Focussing on predictor variables, this means that the key variables affecting the species' distribution across its full potential range have been identified in the model fitted to records of the species in its native range. This is where the concept of **proximal** predictors comes in (*sensu* Austin(2002)). Proximal predictors are those that are most directly and closely linked to the species requirements – i.e. they are functionally relevant. Methods that project environmental relationships to new environments implicitly assume that there is a core set of proximal predictors that are important throughout the whole range of a species.

Proximal predictors contrast with **distal** predictors, which are less directly relevant. Elevation is a well-known example of a distal variable. It has been used as a predictor in many studies largely because it is easy to measure or commonly available as GIS data. However, few organisms respond to elevation *per se*, but rather are sorted along elevation gradients because of associated changes in proximal climatic factors such as temperature, rainfall, solar radiation, and humidity (Austin and Smith 1989). The difficulty in using elevation is that it is only effective as a predictor through its correlations with the more proximal variables. These correlations between distal and proximal variables are imperfect and vary geographically. As a consequence, use of a distal variable like elevation as a predictor will result in models where key relationships are blurred, and predictive power in new regions with different correlation structures is reduced. It is much more straightforward if proximal predictors can be identified and used in modelling.

3.2 Do we know which predictors are proximal, based on theory?

It would clearly be helpful if there were good knowledge of likely proximal predictors for species, or even for taxonomic groups. This is most likely to come from physiological understanding of the requirements of species. Literature searches (Appendix A9.1) reveal useful general information – refer to that appendix for notes on 40 relevant papers published on this topic, and on choice of predictor variables in species modelling. These show in general that (a) there is physiological information about species requirements and tolerances that can inform knowledge in general terms about variables affecting species survival; some examples are provided in Table 1; (b) there are numerous papers comparing use of different variables in species distribution models, but these are simply tests on observed data and goodness of fit of models, and provide no evidence about what is proximal. They do however give clear evidence of the impact on predictions of choice of variables (see below); (c) there are some useful conceptual frameworks for thinking about proximal variables, but as soon as the choice becomes practical, and especially when the

choice is at global scales where there are limited choices of available predictors, the frameworks have some utility, but only in broad general terms, for choosing variables.

A key limitation of the translation of this information into practice is that these proximal variables represent processes operating at different scales, and in real applications for invasive species, the data that would represent the relevant scales may not be available. In addition, a variable may be known to be important for survival and reproduction, but could be expressed in many different ways (annual, seasonal indices etc). For example temperature is widely viewed as a proximal variable, but what aspect of it drives distributions. Is it the maximum, minimum or the mean or a complicated combination of all three? Jackson *et al* (2009) discuss this in detail and give several practical examples of the challenges of dealing with these nuances in species modelling.

Table 1. Examples of proposed variables driving distributions of various groups of organisms.

Group	Proposed variables	Ref for source	Examples of transfer to modelling
Plants	Light, temperature, nutrients, water, CO ₂ , disturbance, pathogens, predators, competitors	Austin (1980)	Austin & Van Niel (2011b), Williams <i>et al.</i> (2012), Mellert <i>et al.</i> (2011b)
	Soil water availability, growing degree days, min winter temp (representing frost).	See refs in Piedallu <i>et al.</i> (2013)	Piedallu <i>et al.</i> (2013)
Insects	Richness explained by energy-water variables (e.g. evapo-transpiration) and their dynamics and seasonality	Diniz-Filho <i>et al.</i> (2010)	See refs in Diniz-Filho <i>et al.</i>
Birds	Temperature	Parmesan <i>et al.</i> (2000)	
Armadillo as e.g. of mammals	Rainfall, temperature (days below freezing)	Parmesan <i>et al.</i> (2000)	

3.3 Spatial scale

Scale comprises both **grain** and **extent**. The extent (or domain) is the geographical area covered by the modelling, and usually reflects the purpose of the analysis. For instance, macroecological and global change studies tend to be continental to global in scope, whereas studies targeting detailed ecological understanding or conservation planning tend toward local to regional extents. Grain usually describes properties of the data (“data resolution”) or analysis - often the predictor variables and their grid cell size or polygon size, but also the spatial accuracy and precision of the species records ((Dungan *et al.* 2002; Tobalske 2002). In invasive species modelling, the grain usually refers to the grid cell (raster) size of the predictor variables.

Grain is the most critical part of scale when thinking about predictor variables and their relationship to processes affecting distributions. Austin and Van Niel (2011b) discuss this, focussing on plants and predictions of distributions of species under changing climates:

“A biologically relevant variable must also have a data resolution that is consistent with the scale at which the ecophysiological processes show greatest variation. Some hierarchical frameworks recommend critical scales for different environmental characteristics but assume the use of coarse- scale data for large areas

and fine-scale data for small areas (e.g. Pearson and Dawson 2003). This is not necessarily appropriate, as local topographic factors may modify the climatic impact, particularly when studies are applied to very large areas (Austin and Van Niel 2011a). Suitable conceptual models of the use of predictors have been presented (Franklin 1995; Guisan and Zimmermann 2000) but are often ignored in climate change studies”.

*And later in same paper, they state: “Plot size [in papers they reviewed and presented in a table] varied from 16 m² to 2500 km². The larger grid cell sizes reflect the interest in climate change, and, importantly, the availability of distribution data at a grid cell size of 50 x 50 km. The assumption that only climate variables are important when the extent of a study is very large leads to the corollary that local environmental heterogeneity can be ignored in large-area studies. Since these assumptions were recognized (Huntley et al. 1995), they do not appear to have been explicitly tested, although see Coudun et al. (2006). Local heterogeneity is important for light (see below) and for soil properties such as nutrients. Soil properties vary with lithology and along topographic gradients from ridge to gully. The magnitude of these local differences in soils will equal or exceed that between 50-km grid cells. Coudun et al. (2006) explicitly tested whether including soil nutrient variables with climate variables improved a model predicting the distribution of the tree *Acer campestre* across the whole of France. It did. Such soil heterogeneity may define local refugia for species, confounding predictions of distribution under climate change.”*

We agree with the authors that there is an important, explicit link between ecophysiological processes affecting species and the implied grain of predictor variables. The problem when modelling species in a biosecurity context is that:

- a) Records of species occurrences used to fit these models are usually from global databases, and these records are rarely all accurately located. Often, spatial locations are only accurate to 1 to 10km of the true location. Even if grids of predictor variables at fine grain were available, this uncertainty in species locations would not allow accurate identification of fine-grained relationships.
- b) Predictor variables must have a global coverage. So far, most global coverages are of climate variables, often at coarse grains. See Table 2 for more information on this.

This means that, despite some good general understanding of processes affecting species, it is often hard to translate it into the choices that need to be made in practice when modelling invasive species.

3.4 Sources of global terrestrial datasets for predictor variables

Table 2 summarises GIS datasets commonly used to source predictor variables for invasive species modelling, plus others available globally but used less often. It focuses on terrestrial data and, for the climate variables, on estimates of current conditions based on long-term climate data. In the published literature the most commonly used variables are the WorldClim and CRU datasets (Table 2). For instance, in March 2015 citations to the WorldClim key reference totalled 5820, compared with about 3000 citations to the 2 papers describing the two versions of the gridded CRU data and 110 citations for CliMond. WorldClim, published in 2005, includes a set of 19 bioclimatic predictors based on temperature and rainfall, at resolutions down to 1km. CRU, published around 2000, includes a more extensive set of variables, but it is coarser grain (~20km, smallest grain). Given that several of the additional variables at CRU and CGIAR (Table 2, e.g. humidity, frost frequency) are likely to be proximally relevant to the distributions of many species, it is surprising – and perhaps indicative of common lack of thought about predictor choice - that so many published papers rely on the WorldClim set. For a comparison of the CliMond and Worldclim variables in their mapped forms, see CEBRA 1402B_comparing mapped variables.pdf, and the similarly named zip file of images.

More data are gradually becoming available; for instance, the availability of global 90m shuttle radar digital elevation datasets makes it likely that terrain-based variables (e.g. topographic position, flow accumulation, valley bottom measures; Gallant and Dowling 2003; Gallant and Wilson 2000) can be estimated globally (John Gallant, pers. comm.). Soils data are also potentially important but as yet not available as a high-

quality data source globally. Existing products tend to be based on sometimes coarse-grained data – e.g. the harmonized world soil database (FAO *et al.* 2012).

Remotely sensed vegetation indices such as NDVI (the normalised difference vegetation index) have also been used in some contexts. Bradley *et al.* (2012) present a useful discussion of the use of vegetation-related indices (including land cover, Table 2) in species distribution modelling. They make the case that including these types of remotely sensed variables might map current distributions well due to potentially tight relationships with distributions of some plants, but might compromise the ability of the models to identify proximal climatic variables affecting the species of interest.

3.5 How has theory been translated into practice, in choice of predictor variables?

Theory-based identification of proximal predictors presents general concepts of the sorts of information that is needed to predict distributions. When faced with modelling a particular species, modellers face the decision of how to use this information to choose from the predictor variables they have identified as available. Modellers do this in more or less structured ways. For instance, Austin and Van Niel (2011b) summarise the approach used in 12 studies of plants, and note strong variation in which of six proximal variables (light, temperature, nutrients, water, disturbance, biota; see section X.2) are represented, and what variables are used to represent them. They find:

*“It is clear that each study has an implicit conceptual model, but there is little consistency between them. No study includes predictors for all six conceptual variables, although the category ‘other predictors’ may provide surrogates for them. The total number of predictors ranges from 5 to 38. The number of predictors used for each conceptual variable varies greatly; for example, for water Pearman *et al.* (2008) used one variable while Coudun *et al.* (2006) used ten. All studies in the table include temperature- and water-related predictors but only six include light. No two studies have identical predictors for temperature or water.”*

Austin and Van Niel (2011b) recommend much more careful and explicit consideration of the proximal variables, how they affect a species and at what scale, and how to link these to predictor variable choice. Whilst this makes much sense and is a good framework for selecting variables, there is then a large gap between what should ideally be done, and what the available data allow. Furthermore, the claim that a conceptual link between *a priori* biological understanding and predictor variables will make models more effective, while plausible, is relatively little tested and remains speculative.

Some researchers have instead started from a very pragmatic viewpoint, and have collected data, proposed reasons for choice of predictor variables, then tested the predictive performance of these various choices. For instance, Barbet-Massin & Jetz (2014) modelled 243 bird species in the USA, aiming to “develop and demonstrate a comprehensive approach for identifying the climatic predictors providing greatest model accuracy”. They estimated their own set of climate variables from the USA’s PRISM climate dataset (<http://prism.oregonstate.edu>), calculating equivalents of the 19 WorldClim variables plus potential evapo-transpiration, growing degree days above 5 degrees and moisture index. They then tested the predictive power of models fitted to subsets of these variables, where the subsets were 6 variables not strongly correlated. Prediction was tested on spatially and temporally distinct testing sets of data. They found that annual potential evapo-transpiration, mean annual temperature and growing degree days produced significantly more accurate SDMs than any other predictors, and that annual precipitation and the moisture index were also useful. This is an example of an approach that is pragmatic and that tests predictive ability on the sort of data available.

See Synes & Osborne (2011) for a useful summary of approaches that have been used in the species modelling literature for selecting variables, and their link (or lack of) to theory – see their section “Creation of variable datasets”. The general concepts are already covered in the sections here, but they provide interesting links to further literature on it.

Table 2. Global GIS data for predictor variables in models.

Name & source	Details of predictors, grain	Comments
WorldClim www.worldclim.org	Monthly grids for min, max and mean temperatures; also monthly precipitation and elevation. Temperature and precip variables also summarised into 19 “bioclimatic” predictors. Grain sizes: 30 arc-seconds (~1km); 2.5, 5 and 10 arc-minutes (i.e. ~ 5, 10 and 20 km)	The interpolation methods used to create WorldClim data and the summaries in the bioclimatic predictors are based on Mike Hutchinson’s approaches for ANUCLIM (Xu and Hutchinson 2013), all rainfall and temperature related. See Hijmans <i>et al.</i> (2005) for details.
CRU http://www.cru.uea.ac.uk/cru/data/hrg/	Mean monthly estimates of: precipitation, wet-day frequency, temperature, diurnal temperature range, relative humidity, sunshine duration, ground frost frequency and wind speed. Grain: 10 arc-minutes (~20km)	This is an update of the first set released in 1999, that had a coarser grain (30’) and fewer weather stations contributing to the interpolation. See New <i>et al.</i> (2002) for details.
Climond www.climond.org	35 “bioclimatic” variables; also mean monthly summaries for daily minimum temperature, daily maximum temperature, monthly precipitation total, daily average radiation. Grain 10 and 30 arc-minutes (~20 and 60km).	These are hybrids of WorldClim and CRU data, combined as described in Kriticos <i>et al.</i> (2012). The 35 predictors are based on Mike Hutchinson’s approaches for ANUCLIM (Xu and Hutchinson 2013), and include radiation and moisture-related indices additional to those in WorldClim. For a comparison of these, mapped, with the worldclim variables, see CEBRA 1402B_comparing mapped variables.pdf, and the similarly named zip file of images.
CGIAR-CSI www.cgiar-csi.org	Aridity, potential evapotranspiration and soil water balance. Grain 30 arc-seconds (~1km)	These are useful grids aligned to the WorldClim variables; surprising that they are not more often used in species modelling
Consensus land cover http://www.earthenv.org	A consensus of 4 inputs, providing estimates of the prevalence of each of 12 landcover types within 1km grid cells (Behrmann projection)	See Tuanmu and Jetz (2014) for methods. This aims to reduce inconsistencies and errors between products, and was tested on fine-grained data.
Elevation		From various sources including the WorldClim and CGIAR sources above. Can be based on models or remotely sensed (shuttle radar) data.

3.7 Does it matter which predictors are chosen?

It is reasonable to question whether the particular predictors chosen from a large candidate set actually affect predictions. One might expect that all are closely related, and the choice of a particular subset has little impact. Several studies show that this choice can have substantial impact, particularly for applications like invasive species modelling, where the model is used to predict to new places. Figure 3 shows the results of Ashcroft *et al.* (2012b) who modelled the emerald furrow bee (*Halictus smaragdulus*) in their native northern hemisphere range and compared use of either all 19 WorldClim variables or a subset of 4 (variables 1, 5, 6 and 12, Appendix 9.1) or 2 (variables 1 and 12 – annual temperature and annual rainfall). Predictions into Australia vary substantially depending on the selected set (Figure 4) despite the similar performance of the SDMs in the native range (Figure 3).

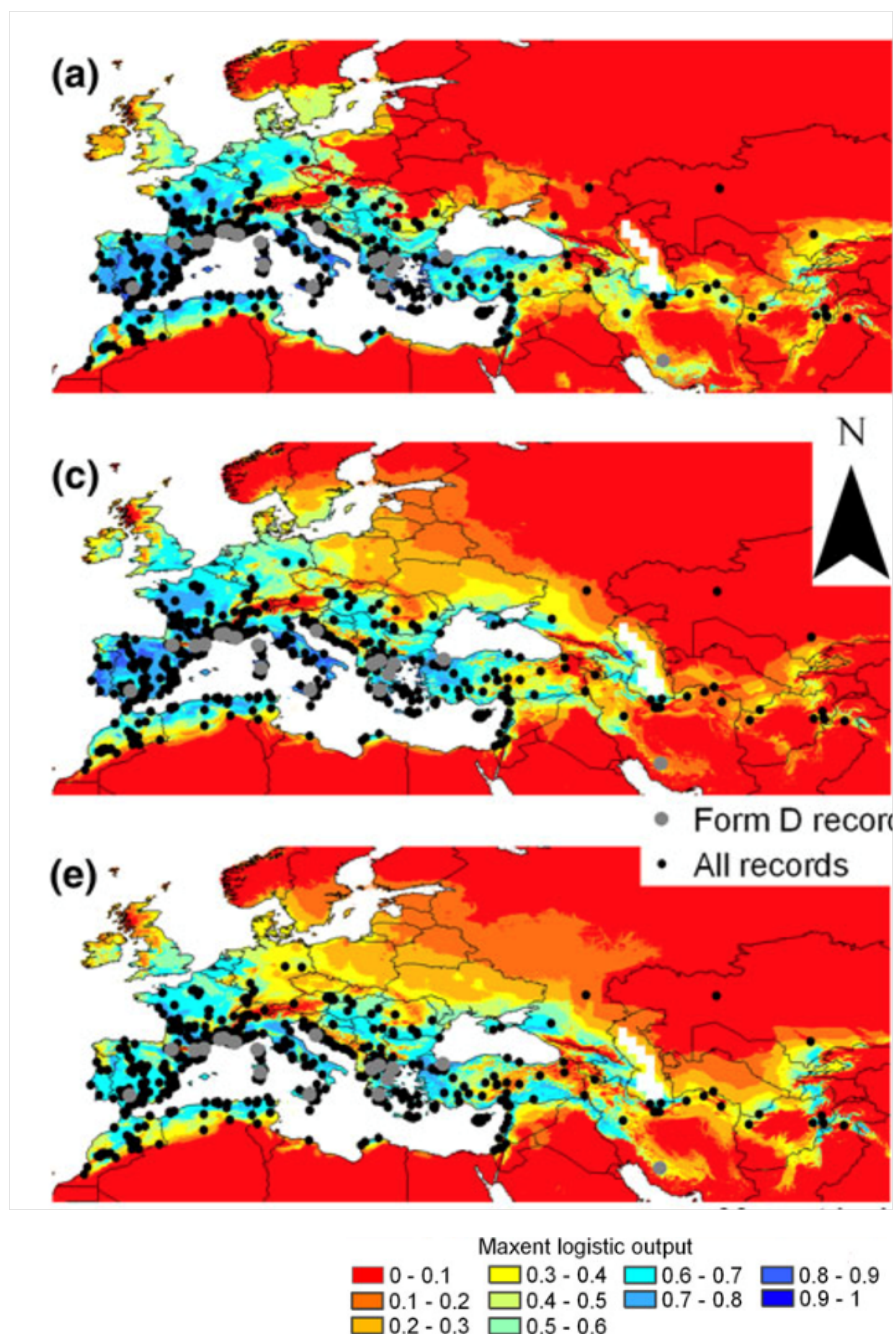


Figure 3. Known locations for *Halictus smaragdulus* (black & grey dots), and Maxent predictions in native range for models fitted to (a) 19 (c) 4 and (e) 2 climate variables (see text).

3.8 Dealing with relationships between predictor variables

Within datasets of climatic variables there are often several highly correlated variables. For example, the set of 19 WorldClim variables (Table 1) includes temperature and rainfall-related indices (Appendix 9.2), and in any particular region several are often highly correlated. Common practice in dealing with these varies, and includes:

- do nothing; include all variables of interest;
- choose a subset based on ecological knowledge, with or without consideration of correlations between variables;
- choose a subset based on explanation of each variable, singly, in a model; or on change in explanation when dropping one (summarised in Synes and Osborne 2011)
- use one of several techniques including pairwise tests of Pearson correlations, principal component analyses (Dormann *et al.* 2013) to test the relationships between variables and select a subset based on the results. See also the method in Williams *et al.* (2012), that estimates dissimilarity between rasters of variables using a Gower-style metric (see their Section 3.3).

While in our opinion the impact of correlated variables is overstated in the literature, a practical issue remains. The problem with making no initial selection between candidate predictors is that the species data sets usually available for invasive species modelling are often small to moderate in size (5-200 observations). So even if there were reasons to consider all variables there is limited scope to reliably fit model parameters to large numbers of variables. Many argue that it is best to make initial selections based on ecological reasoning and understanding of the relationships within the data. In practice this remains complicated given the range of representations of conceptual concepts such as temperature and moisture availability.

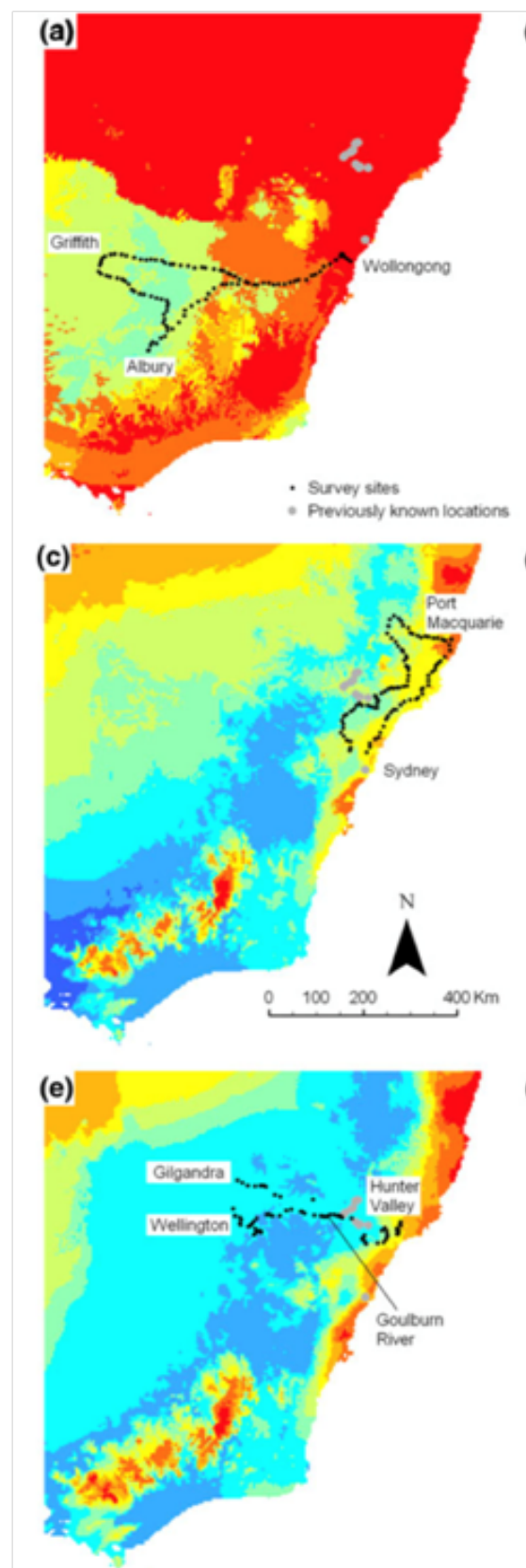


Figure 4. Projections of Maxent model for *Halictus smaragdulus* into south east Australia from the models shown in Fig. 3.

3.9 Discussion

The literature is surprisingly short of concrete recommendations regarding proximal variables. While the concept is often invoked to support particular choices of variables, the proof of their validity is typically by assertion. The difficulty is that correlation can mimic causation, and that incomplete information can mask causation.

Thus with organisms that are well studied by competent physiologists it may be possible to identify key variables that are broadly predictive of a species distribution. It would thus be the recommended approach, noting that it is still an emerging area of science. But for less well known organisms it is more difficult to make general recommendations. There is a universal view that both temperature regimes and patterns of moisture availability have major causal impacts on species distributions. The challenge is that these relationships may be complex and choosing a particular representation on an expert basis may be challenging. We consider an alternative in the next chapter.

4 Outline of modelling process

Chapter 5 introduces the case studies and details of data selection and tests regarding the various steps of model building, but here we outline the general modelling process.

Initial models fitted to native range data

The models fitted to records in the native range were fitted using a Generalized Additive Model (GAM) implementation of a Point Process Model (PPM) for presence-only data (Renner *et al.* 2015; Renner and Warton 2013; details in Chapter 5).

Development of climate envelopes

The previous project (CEBRA Project 1302) identified that probability based projections would typically perform poorly as the underlying populations were different. In other words, if a model is fitted to say presence-absence data in the native range, the probabilities will be correctly calibrated in the native range, but they may not apply well to the invaded range – the proportion of presences observed may be significantly different and predictions about places that are more or less suitable might be unreliable. This is explored later in this report. Whilst we recognise that there is much to be explored around the question of whether there is *some* useful information in the continuous predictions from fitted models (e.g., whether relative values are informative), we take the route here of trying to build robust binary predictions. Therefore as an alternative to probability based predictions we consider the use of climate envelopes, to be applied to variables selected after the first step of model fitting in the native range. Climate envelopes involve a binary or ordinal classification system. Usually, the environment is classified as either suitable or unsuitable. They can be interpreted as a regression technique in the sense that they determine habitat suitability on the basis of the environment. But they are conservative in the sense that locations within the envelope are not differentiated in terms of suitability..

There are a number of approaches to determining climate envelopes in the literature. The Bioclim model involved developing a rectilinear region in environmental space based on analysis of the observed data's position in that space (Longmore *et al.* 1986). Later authors sought to restrict the envelope more closely to the data. Instead of analysis of individual variables one at a time they considered using distance based methods (DOMAIN, CLIMATCH) to construct local envelopes defined by the presence points. These methods define some neighbourhood of each presence point to be associated with suitable environment. This approach allows interactions between variables to be incorporated.

While the consideration of interactions is useful there are a number of issues with this approach. In particular, there are challenges in creating these envelopes when the data is multi-dimensional. In this case a phenomena termed the “curse of dimensionality” affects our ability to estimate envelopes in high dimensions. The issue in this case is the data become increasingly sparse as the dimensions, in this case the number of environmental variables, increase. As an example we simulated 1000 data points from the uniform distribution over [0, 1] for each of 10 dimensions (in other words, sampling from a 10 dimensional distribution with uniform and independent margins). We then calculated the mean distance to the nearest neighbour in the data set for dimensions one to ten. This is shown in Figure 5. Note that the distance increases as the dimension increases.

What this means for the construction of environmental envelopes is as follows. If we fix the distance that we associate with presence for each point we end up with a climatic envelope that has “holes” in it as the number of environmental variables increases. This is because the presence points become further apart as the data become sparser in high dimensions. If we adapt the local distance two problems emerge. First we are not sure how to effectively calibrate this. Second, increasing the distance leads to greater smearing of the distribution across the edges of the envelope. This means we increase the region that we assign as

suitable when it is in fact not. In high dimensions there are more “edges”. So we have more edges to estimate and a greater propensity to over-predict.

We aim to test models for situations where information about the physiology of the organism is limited. We take the approach of testing smaller models for which the climate envelope can be robustly estimated. We fit low dimensional (2-3 variable) models. In lower dimensions we are able to detect holes in the envelope more easily. These envelopes will also be broader as they are not “specialised” to large numbers of variables. They will also not overfit the native distribution so will hopefully be more conservative in their projections.

Projection methods

Projections are predictions from the fitted model to the whole world. We explored three methods of performing global projections of each model. The choice of projection method depends on the structure of the fitted model. The first method for projection is appropriate for point process models and involves making a prediction of the Poisson intensity at each location across the globe based on the fitted model. This projection method results in a gradient of low to high intensity but does not produce a probability. The prediction output is instead a measure of the predicted population intensity per unit area and is thus relatively meaningless in an absolute sense. The projection does give a relative abundance prediction though.

The second global prediction method explored is the use of a minimum bounding box in environmental space. This is an envelope method. Essentially all presence points in the native range are plotted in environmental space using the chosen environmental variables. The minimum bounding box is then constructed by drawing a box in environmental space, taking the minimum and maximum values of each of the chosen variables as the endpoints of the box. The projection is then made to the globe by determining if each geographic location is inside (predicted presence) or outside (predicted absence) this box. The resulting global prediction depicts all places on the globe with environmental conditions within the range of where the species is found in its native range.

The third global prediction method is a restriction of the second method. Instead of a bounding box being constructed in environmental space, an alpha-hull is constructed. The alpha-hull is a shape constructed using a parameter alpha. Alpha is the diameter of a sphere in environmental space. As alpha gets smaller, the alpha-hull shrinks to the limit where it includes only the presence points. As alpha get larger, the alpha-hull increases to the limit where it coincides with the minimum bounding box. The value of alpha we have chosen is the value which creates the minimum sized box shape that includes all presence points and does not have any internal holes. These holes would be difficult to justify physiologically. This could be termed the minimum bounding region. Global prediction then proceeds in the same manner as method two, where all geographic points are determined to be inside or outside the alpha-hull.

Convex hulls are an alternative, less computation intensive approach to using alpha-hulls. A convex hull is a compromise between the use of a minimum bounding box and an alpha-hull. The alpha-hull is smaller in most every case, except when the alpha-hull is also convex. Convex hulls and bounding boxes suffer from the problem that they will generate biased estimates of a species tolerance of environmental conditions, even with perfectly accurate and abundant data, if the shape of a species true environmental envelope is not convex. Alpha hulls will converge on the true underlying envelope, as data improve.

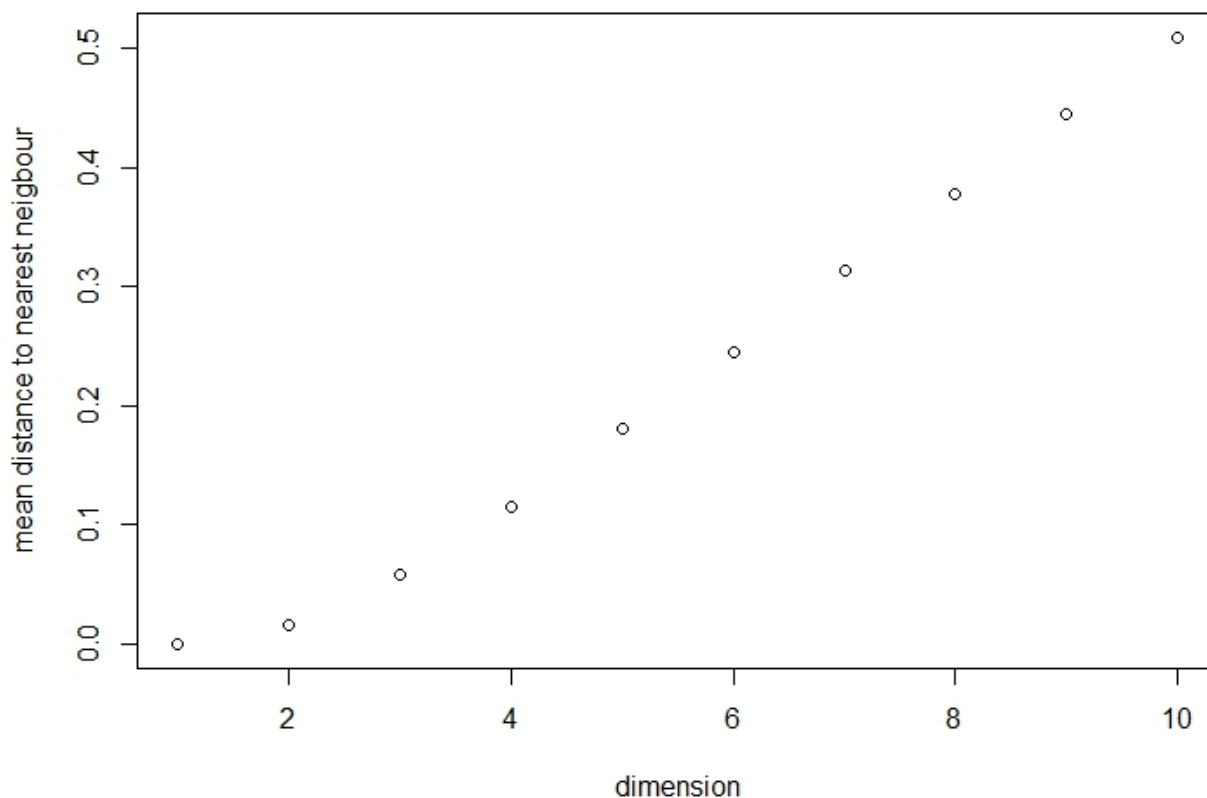


Figure 5 Effect of dimension on the distance to the nearest neighbour for a fixed sample size.

Novel environments

When statistical models like Maxent or GLMs or GAMs are fitted (trained) to species data, predictions can be made globally. However environments may exist globally that are outside the range of the data used to fit the models. These are often referred to as “novel” environments. It is not generally a good idea to use statistical models to predict to novel environments, because the model is ignorant of the species’ response to these new environments. Further, the model is usually also not structured in a way that can be guaranteed to predict sensibly – e.g. the model might predict higher suitabilities at higher temperatures, and not be controlled to predict zero suitability above some threshold. In recognition of these problems, researchers have developed various techniques for quantifying which environments are novel. The sampled environments are the combined set of presence records and background samples. Novelty might be measured, for instance, as:

- those environments outside the range (minimum and maximum) of the sample, as estimated for each environmental predictor in the model, or
- those environments outside a convex hull drawn around the sampled environmental space.

Methods proposed include:

- Elith *et al.* (2010) (“Multivariate Environmental Suitability Surfaces”, or MESS maps), which is implemented in Maxent and available as R code in the package “dismo” (Hijmans *et al.* 2015);
- Zurell *et al.* (2012) – testing new combinations of environments
- Owens *et al.* (2013) – an extension of MESS maps

- Mesgaran *et al.* (2014) – uses Mahalanobis distance; supplied as a stand-alone package “ExDet” (<https://www.climond.org/ExDet.aspx>); example R code given elsewhere (<https://pvanb.wordpress.com/2014/05/13/a-new-method-and-tool-exdet-to-evaluate-novelty-environmental-conditions/>)

These estimates of novel environments can then be used to mask (i.e. to obscure) predictions in novel space, or at least warn about where the model is extrapolating.

Those predictions that use hulls to define the occupied environments and then map those environments globally will not be extrapolating into novel space. In contrast, the predictions from the fitted GAMs may extend into novel space.

5 Empirical identification of proximal variables.

In this section we document our investigation into approaches to identifying proximal variables.

5.1 Protocol development for model fitting

General

We explored general approaches to the development of a protocol for fitting SDMs. The following issues were identified when developing the approaches to be applied to the case studies:

- Fewer variables (e.g. max. of 3) will more likely overestimate the potential distribution, as the model will not overspecialised.
- Proximal variables are strongly predictive. Because they should predict strongly to identify proximal variables we should pick variables that have strong predictive performance.
- Variables on their own aren't "proximal". The combinations of variables provide a biologically plausible and effective model.
- Improved predictive performance may be possible with variables that interact more directly with a species during its lifecycle (e.g. micro-climate data), or are "process" based.
- Choice of background data has a significant impact on model outcomes.

Our approach can be thought of as an application of the current correlative modelling approaches in the literature, though with severe restriction on the number of variables (restricted to 2-3 only). A further requirement was that each model could only contain one "temperature" type variable and one "precipitation" type variable (see below). The rationale is that such an approach should avoid overfitting in the training range (typically but not exclusively the native range), and hence generate better robustness when projecting.

Predictor variables

Variables were chosen from the original BIOCLIM bioclimatic variables (BIO1 – BIO19) as generated for the globe in CLIMOND (Kriticos *et al.* 2012). These variables and their descriptions are presented in Table 3.

Kearney *et al.* (2014b) created physiologically motivated climatic variables that provide global estimates of hourly microclimate. The name "microclimate" emphasises climate near the ground, in contrast to the commonly used SDM climatic predictors which are long-term average estimates of climate at standardised heights above ground. Since many organisms actually experience near-ground climates, microclimate is likely more proximal and hence informative. These variables are estimated through mechanistic microclimate models which include routines for hourly calculations of solar radiation intensities, above-ground profiles of air temperature, wind velocity and relative humidity, and soil temperature profiles. These models can be applied to base gridded climate and soil data of various temporal and spatial resolutions. In this project the approach of Kearney *et al.* (2014b) was – for the first time - used to generate MICROCLIM variables on the same basis as the BIOCLIM temperature variables. That is, the BIO1 – BIO7 variables were reproduced using soil temperature at the surface and various depths.

We also explored the possibility of using more process oriented components than considered so far, though retaining the restriction on model complexity. That is, we considered a temperature-based index plus a moisture-based index. The idea is to use similar structural elements to CLIMEX (Sutherst and Maywald 1985) but with a more structured approach to fitting. This requires developing fast code (C++ interfacing with R) that can generate CLIMEX variables as described by Sutherst & Maywald (1985) quickly. We developed this based on monthly temperature and rainfall data variables available through the CLIMOND

dataset (Kriticos *et al.* 2012). Issues arose with documentation of the empirical relationship between relative humidity and evapo-transpiration so the CGIAR-CSI (Chapter 3, Table 2) evapo-transpiration data was used instead.

An example of how the new CLIMEX derived variables may or may not influence results is shown in Figure 6. Whilst there are areas where the relationship is uncorrelated, there is clearly a broad correlation between a somewhat complex index calculated across the entire year and a single monthly measure.

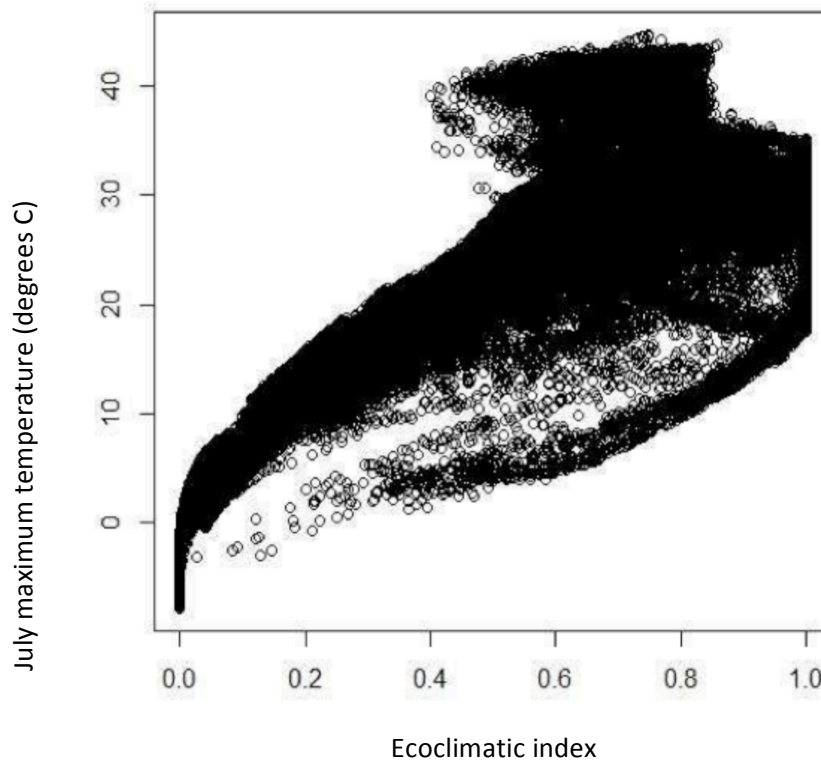


Figure 6. Relationship between July maximum temperature and the ecoclimatic temperature index

The ecoclimatic index is based on this set of temperatures: 0, 15, 35 and 45 degrees for t0 to t3 respectively (for details see Sutherst and Maywald 1985)

Models

As mentioned in chapter 4, the first models were fitted using a Generalized Additive Model (GAM) implementation of a Point Process Model (PPM) for presence-only data (Renner *et al.* 2015; Renner and Warton 2013). Smooth terms were restricted to 2nd order polynomials at the most. Terms were either fitted additively or jointly. Joint fits can be achieved in the *mgcv* package in R by fitting a smooth surface in two dimensions over the chosen variables rather than two one-dimensional fits, one for each variable. In each case the best model was chosen based on proportion of explained deviance. For the BIOCLIM candidate models, this yielded 11 temperature x 8 precipitation x 2 fits = 176 models.

MICROCLIM candidate models were restricted to 1 MICROCLIM temperature variable + 1 BIOCLIM moisture variable. This resulted in 28 temperature (1st 7 BIOCLIM variables at soil depths of 0, 5, 50 and 100 cm) X 8 precipitation x 2 fits = 448 Models.

These GAMs can be directly used to predict relative intensities, as explained earlier. As mentioned in Chapter 4, we also explored the use of climate envelopes as a means of predicting suitability on a binary scale. To achieve this, we took the two variables from the best fitting GAM, and constructed an alpha-hull surrounding the presence points in environmental space. The choice of alpha was such that a minimum bounding envelope was constructed that contained all presence points (not necessarily convex and without holes). To construct a global suitability prediction, all points were classified as either within (a predicted

suitability) or outside (predicted non-suitability) the hull. We also constructed a minimum bounding box in environmental space to construct a similar global suitability prediction. The rationale for such an approach is that these environmental conditions are presences in the native range, so similar environmental conditions should lead to suitable habitat in new locations.

Backgrounds

In order to fit presence-only data in point process models, background points (or quadrature points, Renner et al 2015) need to be selected. The user needs to select the number of points and the extent over which the points are selected. We explored three levels of background extent. The 'local' background was a box in geographic space (excluding ocean) that enclosed all presence points in the native range. A 'continental' background was constructed similarly, restricted within continental borders (except for those that span continents, e.g. species occurring in Central America). The boundaries would be considered "generous", and were typically several multiples larger than the minimum box containing the observed presences, though restricted by the size of the continent in question (see examples in appendices in the separate document: CEBRA 1402B_appendices.pdf). Global backgrounds are as stated, consisting of points over the entire globe. In all cases, 100 000 background points were chosen.

Expert based assessment

For each case study we developed an expert based assessment of possible proximal variables. This was done by one author (Elith) reviewing the literature on the particular species.

Resolution

All analyses were undertaken on datasets at a 10' resolution (~ 19km at the equator).

Table 3. BIOCLIM bioclimatic variables and their description.

Variable	Description
BIO1	Annual Mean Temperature
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3	Isothermality (BIO2/BIO7) (* 100)
BIO4	Temperature Seasonality (standard deviation *100)
BIO5	Max Temperature of Warmest Month
BIO6	Min Temperature of Coldest Month
BIO7	Temperature Annual Range (BIO5-BIO6)
BIO8	Mean Temperature of Wettest Quarter
BIO9	Mean Temperature of Driest Quarter
BIO10	Mean Temperature of Warmest Quarter
BIO11	Mean Temperature of Coldest Quarter
BIO12	Annual Precipitation
BIO13	Precipitation of Wettest Month
BIO14	Precipitation of Driest Month
BIO15	Precipitation Seasonality (Coefficient of Variation)
BIO16	Precipitation of Wettest Quarter
BIO17	Precipitation of Driest Quarter
BIO18	Precipitation of Warmest Quarter
BIO19	Precipitation of Coldest Quarter

5.2 Case studies

We evaluated the approaches using five species, namely the fire ant (*Solenopsis invicta*), Asian gypsy moth (*Lymantria dispar*), Oriental fruit fly (*Bactrocera dorsalis complex*), myrtle rust (*Puccinia psidii s.l.*), and cane toads (*Rhinella marina*). For each species we describe previous modelling efforts from the literature, and compare and contrast our current approaches.

5.3 Protocol development for model fitting – Case Studies

In the ensuing case studies, we first present a brief review of modelling efforts to date from the literature. We then explore, for various backgrounds, a range of different possible approaches:

- Fit GAMs using one of the choices of variables (BIOCLIM, MICROCLIM, CLIMEX, and 'Expert') and backgrounds (local, continental, global).
- For BIOCLIM, MICROCLIM, and CLIMEX variable options, choose the best fitting GAM based on residual model deviance.
- Make a global prediction of the chosen model using the predicted Poisson intensity of the GAM.
- Make a global prediction of the chosen model using an alpha-hull.
- Make a global prediction of the chosen model using a bounding box.

Global predictions are also constructed using a GAM derived from either the first 19 BIOCLIM or all 35 BIOCLIM variables, one of the defined backgrounds, and one of the Poisson intensity, alpha-hull, or bounding box methods.

Full details of all best fitting models are presented in Appendices I – V.

In the following sections we present a selection of the models we produced to showcase those that appear to be predicting the best, and alternatives that demonstrate various issues in the model fitting and prediction routines.

5.3.1 FIRE ANT (SOLENOPSIS INVICTA)

Background

Fire ants (*Solenopsis invicta*) are native to southern Brazil, Paraguay, Uruguay, Bolivia and north-eastern Argentina. They have invaded the US, first in Alabama in the 1930’s, then spreading throughout the south-eastern US and into Texas, New Mexico and California, with northern limits around Maryland and Delaware. They have also recently invaded several Caribbean islands. In Australia and New Zealand they have been found in south-east Queensland, at Yarwun in central Queensland, and at Port Botany in NSW; also at Auckland airport and the Port of Napier, NZ (Commonwealth of Australia 2015; DAFF 2015; Fitzpatrick *et al.* 2007; Sutherst and Maywald 2005)

Fire ants are a serious pest, with social, environmental and economic impacts (DAFF 2015). Their ecology is reasonably well known (details at DAFF 2015). Efforts to model their potential distribution globally include the following (Table 4), and see CSIRO (2015) for a reference list of other relevant publications.

Table 4. A selection of species distribution models from the literature for fire ants (*Solenopsis invicta*)

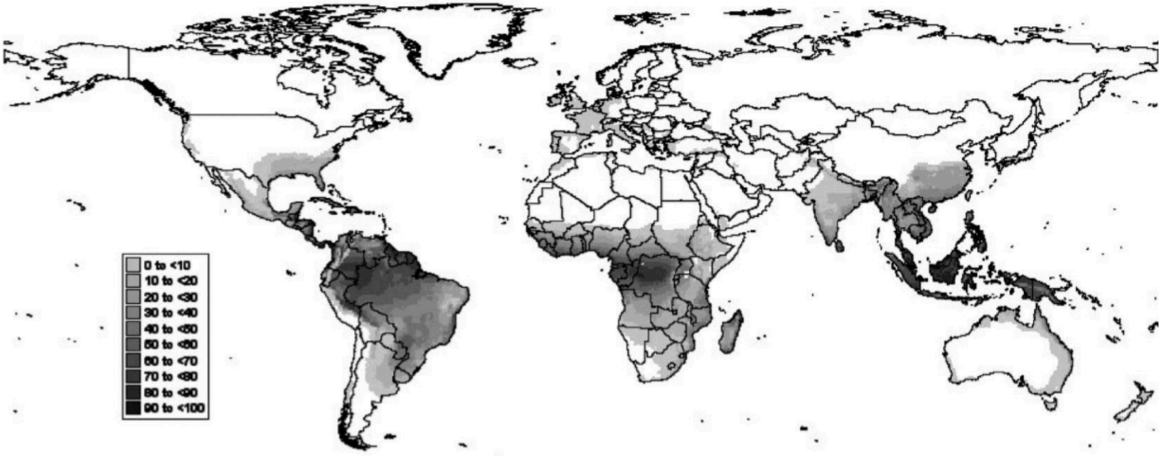
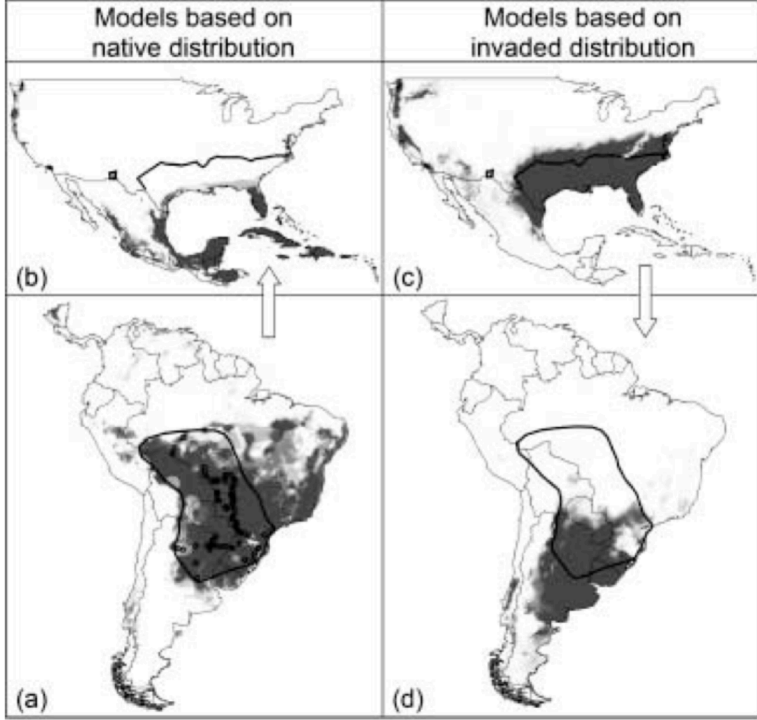

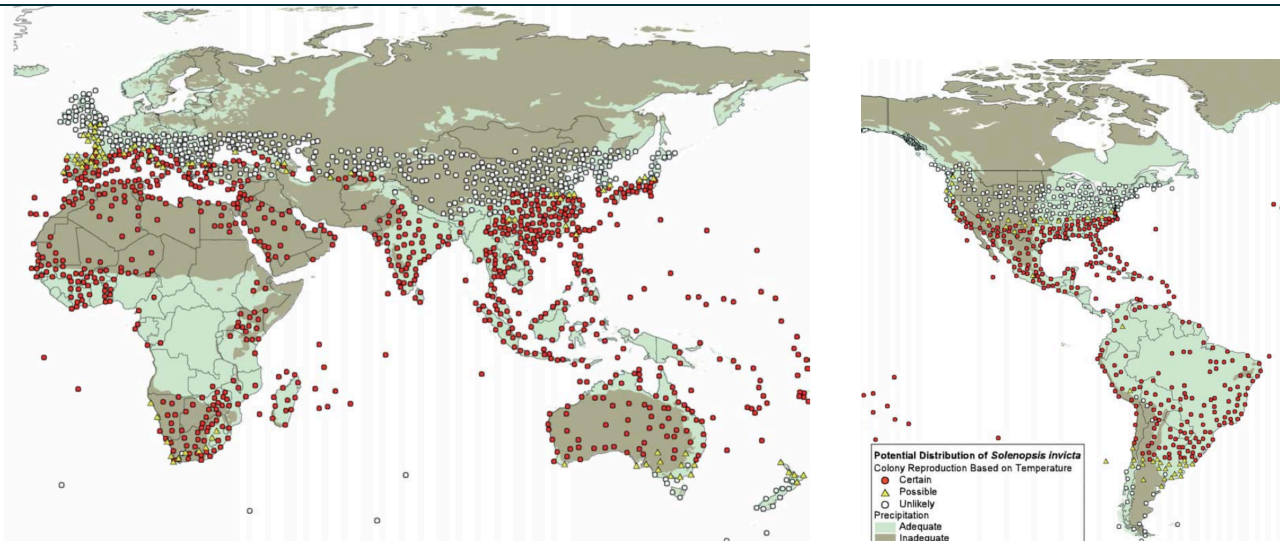
Source	Model details	Mapped predictions
Sutherst & Maywald (2005)	CLIMEX – fitted the model to data in the US, and projected world-wide. Explored likely effects of irrigation on projected distributions. Fit parameters for soil moisture, temperature, and stress (dry, wet, cold, heat), and included constraints based on degree-days to complete generation and to produce alates. They comment in discussion that the fire ant distribution in S America (native range) is riverine and not climate-limited – mentioned role of soil disturbance and inundation.	 <p>The figure is a world map showing the potential global distribution of the fire ant (<i>Solenopsis invicta</i>) based on the CLIMEX ecoclimatic index. The map is divided into regions, with shading indicating different levels of suitability. A legend in the bottom left corner provides the following ranges for the index: 0 to <10, 10 to <20, 20 to <30, 30 to <40, 40 to <50, 50 to <60, 60 to <70, 70 to <80, 80 to <90, and 90 to <100. The highest suitability (darkest shading) is concentrated in South America, particularly in the southern and eastern parts, and in parts of Africa and Asia. Lower suitability is shown in North America, Europe, and Australia.</p>

Fig. 11. Potential global distribution of the fire ant under natural rainfall as estimated by the CLIMEX ecoclimatic index, 21

Source	Model details	Mapped predictions
Fitzpatrick <i>et al.</i> (2007)	GARP. Study of what “reciprocal modelling” (native vs invaded range fits, projected elsewhere) suggests. Predictors: used elevation and 11 Worldclim variables at 10' resolution (bio1,2,3,4,5,6,7,12,13,14,15). Unclear what background was used, being reciprocal modelling perhaps it was continent-wide (S America vs USA). See pictures – showed that native range records fail to predict USA distribution and vice versa). Quite a long and interesting discussion of what all this means.	 <p>The figure consists of four maps arranged in a 2x2 grid. The top row shows maps of North America, and the bottom row shows maps of South America. The left column is labeled 'Models based on native distribution' and the right column is labeled 'Models based on invaded distribution'. Map (a) shows the native distribution in South America with a dark shaded region in the southeast. Map (b) shows the predicted distribution in North America based on the native distribution, with a dark shaded region in the central and eastern parts. Map (c) shows the native distribution in North America with a dark shaded region in the central and eastern parts. Map (d) shows the predicted distribution in South America based on the invaded distribution, with a dark shaded region in the southeast. Arrows indicate the direction of the reciprocal modelling process: from (a) to (b) and from (c) to (d).</p>
Ward (2009)	Specifically focussing on NZ and using 4 “modelling” approaches – in order, left to right: DOMAIN on 19 WorldClim variables (native and invaded range records), climate matching on 4 climate variables, growing degree days (mapping them from published info, using LENZ data) and “foraging activity” which is based on 10cm soil temp data, also from LENZ. Made a consensus map)	 <p>The figure shows five maps of New Zealand, arranged horizontally. Each map represents a different modelling approach: DOMAIN on 19 WorldClim variables, climate matching on 4 climate variables, growing degree days, and foraging activity. The maps show the predicted distribution of the species in New Zealand, with the distribution becoming more detailed and accurate from left to right.</p>

Source	Model details	Mapped predictions
Morrison <i>et al.</i> (2004)	A “dynamic, ecophysiological model” of colony growth based on min and max temps; superimposed precipitation data to identify too-dry places. The model assumes soil temp is key factor affecting colony survival, and is based on estimates of how many alates are produced per female are calculated. Threshold number necessary from USA northern range (coldest area known suitable). Used daily temp data. Threshold of 510mm rainfall selected.	

For our models, we obtained presence records from Matt Fitzpatrick Pers. Comm.), who collated records from native and invaded ranges for his published modelling, as described by him: “Native and introduced distribution data sets consisted of presence data only. We collected 74 native range occurrences of fire ants within South America from primary literature. For invaded range occurrence data, we used latitude–longitude centre points of only those US counties under ‘entire county quarantine’ by the US Department of Agriculture, Animal and Plant Health Inspection Service, which constituted 741 counties in 2004” (Fitzpatrick *et al.* 2007; references provided in quotation but omitted here). The USA county data included information of first recorded infestation.

We restricted ourselves to using the 74 native range recorded presences in our GAM modelling, for selecting the predictor variables. Alpha and minimum bounding box projections were done once using the 74 presence points, and once using all available presence points (native and US).

Results, Fit-based:

The results for the best fitting BIOCLIM temperature (BIO3) and precipitation (BIO18) model for the distribution of the fire ant *Solenopsis invicta* are shown in Figure 7, based on observed locations in the native range of South America and a continental-scale background. These use the model to predict relative intensities worldwide. As with the previous modelling effort of Fitzpatrick *et al.* 2007, the model fails to predict (omission errors) the observed distribution in the south-eastern United States (Figure 7). At the Australia/New Zealand scale the projection is considered mixed, being misleading in places. The distribution in Australia is plausible based on observed incursions (Brisbane and Gladstone), though the predicted suitability of cool moist mountainous regions in New Zealand and parts of Tasmania doesn’t accord with what we think we know (Figure 8). In combination with its inability to predict the invaded US range, we would not consider this

model reliable for projection. Allowing the model to be chosen from the best performing MICROCLIM temperature variable and BIOCLIM precipitation variable results in little improvement (Figure 9).

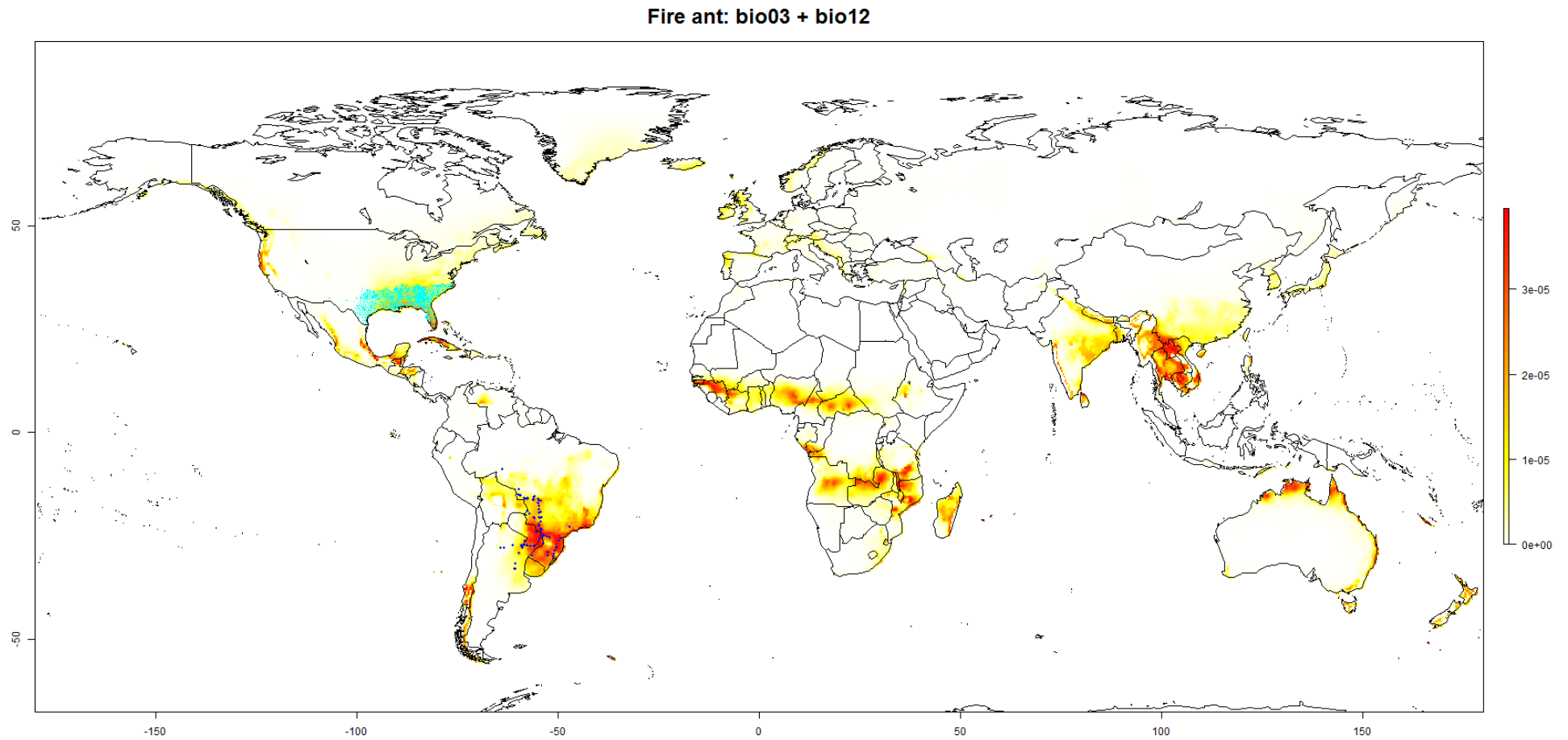


Figure 7. Best fitting GAM using BIOCLIM temperature (BIO3) and precipitation (BIO18) for fire ants (*Solenopsis invicta*) with continental scale background. Blue crosses in South America denote occurrences. Aqua dots in the USA denote the invaded range (not used in model development).

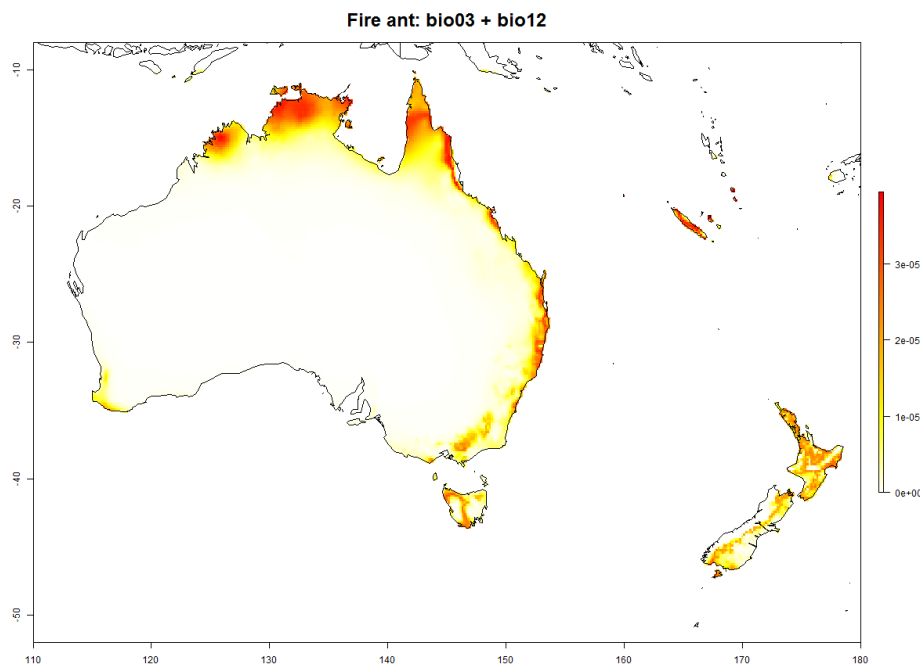


Figure 8. Projection of best fitting GAM using BIOCLIM temperature (BIO3) and precipitation (BIO18) for red imported fire ant to Australia and New Zealand. Background is at continental scale.

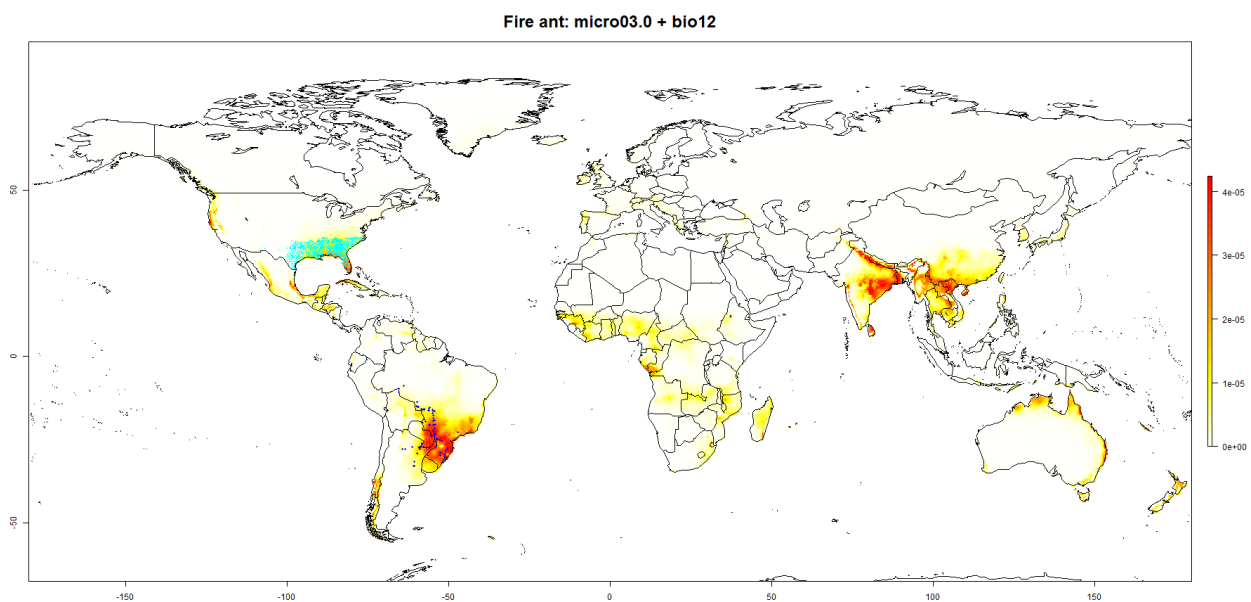


Figure 9. Best fitting GAM using MICROCLIM temperature (O3.0) and precipitation (BIO18) for fire ants (*Solenopsis invicta*) with continental scale background. Blue crosses in South America denote occurrences. Blue dots in the USA denote the invaded range (not used in model development).

The equivalent GAM using a local background performs poorly, particularly in the native range, though it performs better in the southern USA (Figure 10). The likely reason is the over-fitting to local features of variables. This was apparent for many of the modelling approaches (see Appendix I in CEBRA 1402B_appendices.pdf). A similar result was found for the global background (see Appendix II, V in CEBRA 1402B_appendices.pdf). Neither global nor local background is considered further.

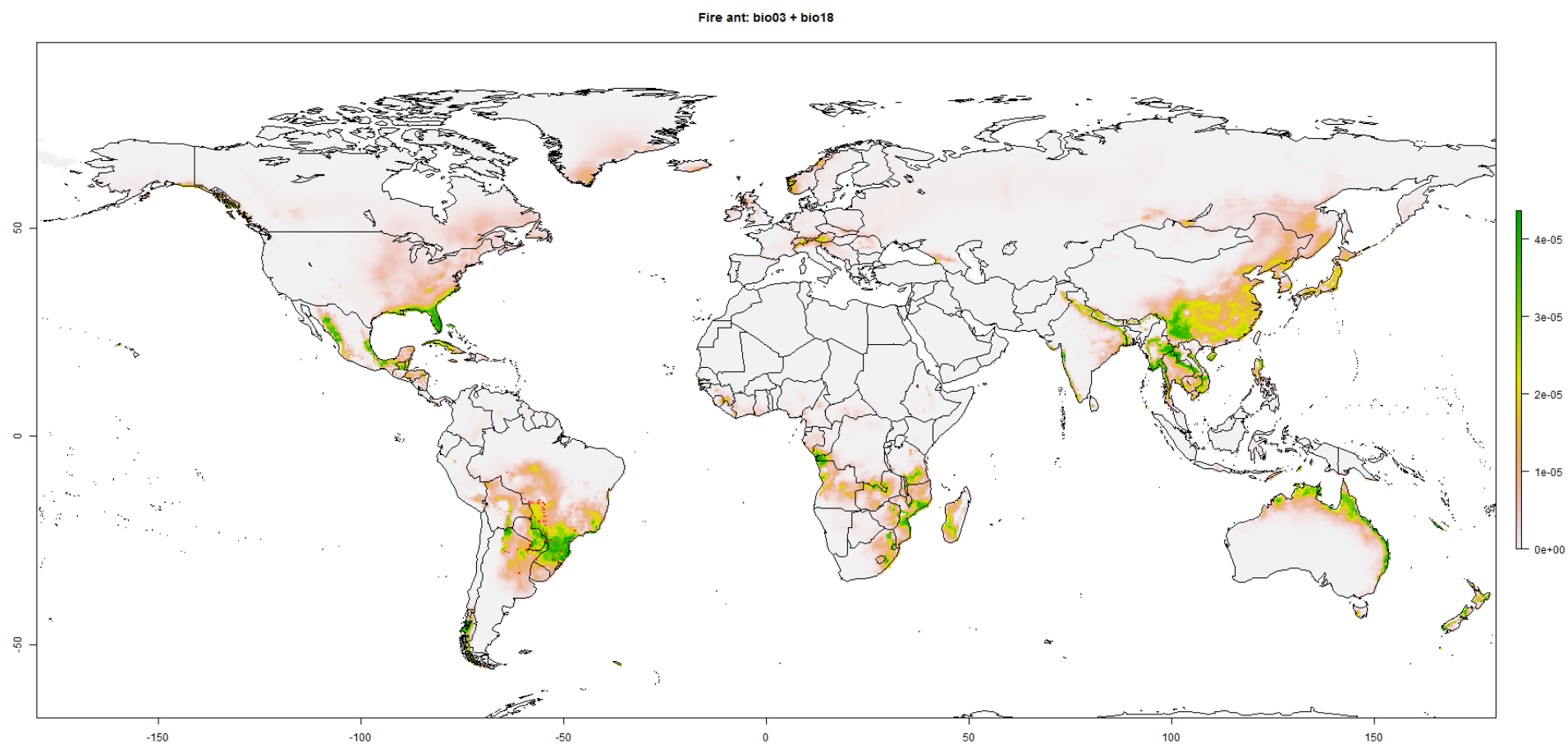


Figure 10. Projection of best fitting GAM using BIOCLIM temperature (BIO3) and precipitation (BIO18) for red imported fire ant to Australia and New Zealand. Background is at local scale.

Neither does using expert chosen variables appear robust to whatever the inherent problems are in the fire ant data (Figure 11).

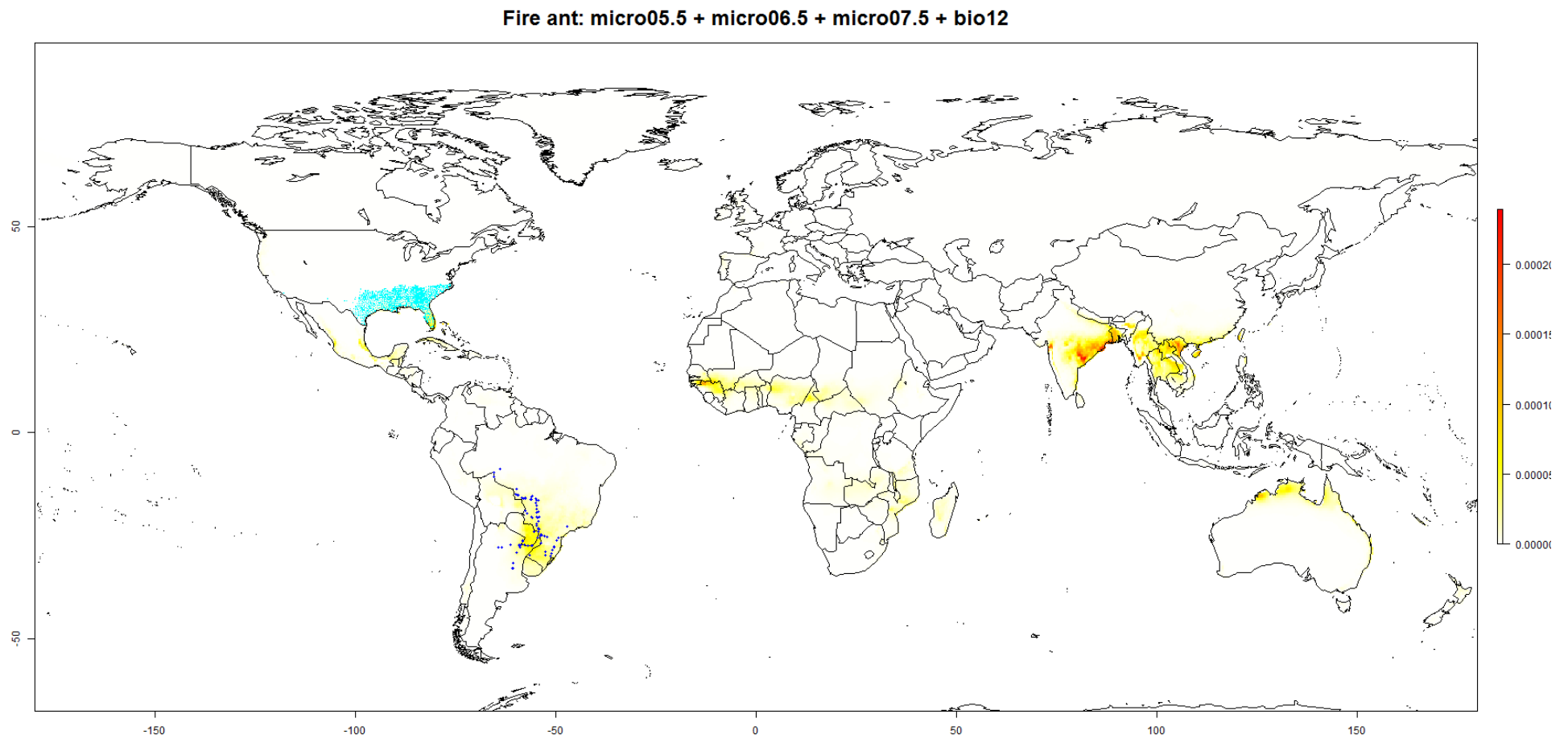


Figure 11. Projection of best fitting GAM using expert chosen variables for red imported fire ant to Australia and New Zealand. Background is continental.

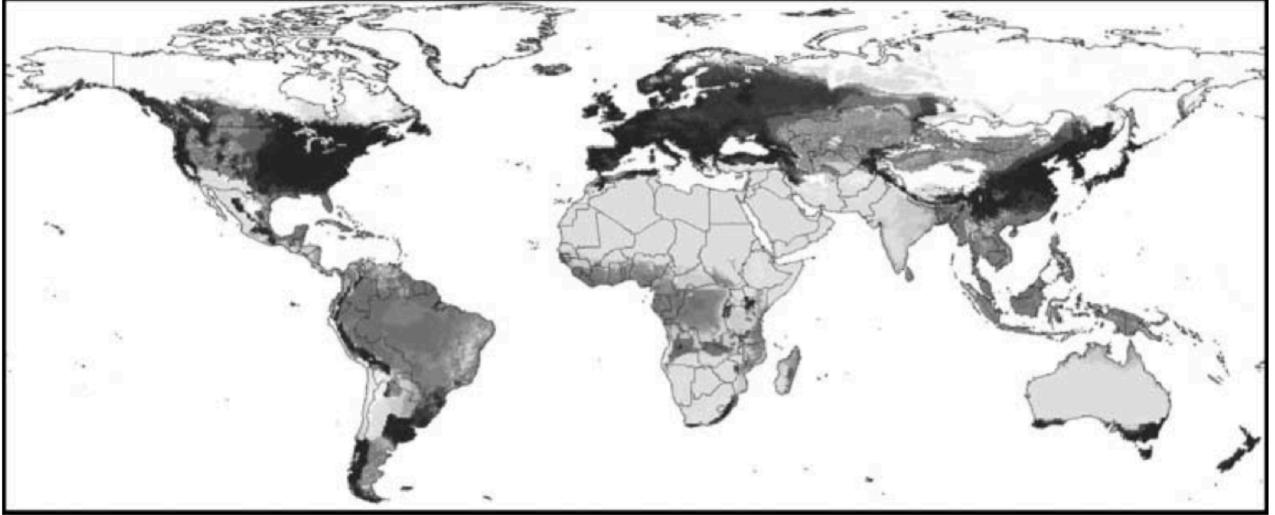
Clearly no modelling approach performs satisfactorily in predicting the invaded range in North America. Sutherst & Maywald (2005) considered that unquantified biotic interactions precluded using the South American native range for estimating model values. Such inference was made in the knowledge of the invaded range, however, such knowledge will commonly be missing. Alternatively, it could be the quality of data from the native range may be poor (e.g. only collected along rivers or roads). Again, such metadata on data quality will often be missing.

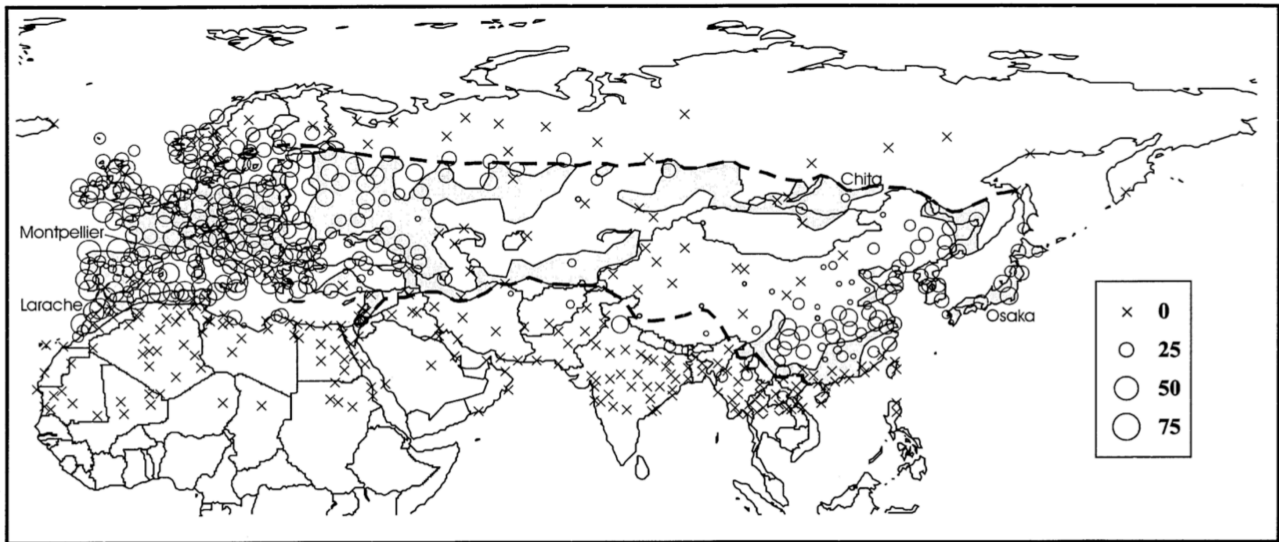
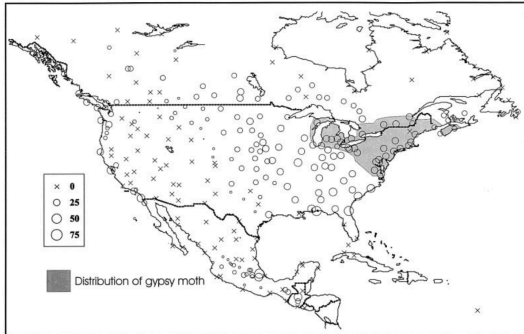
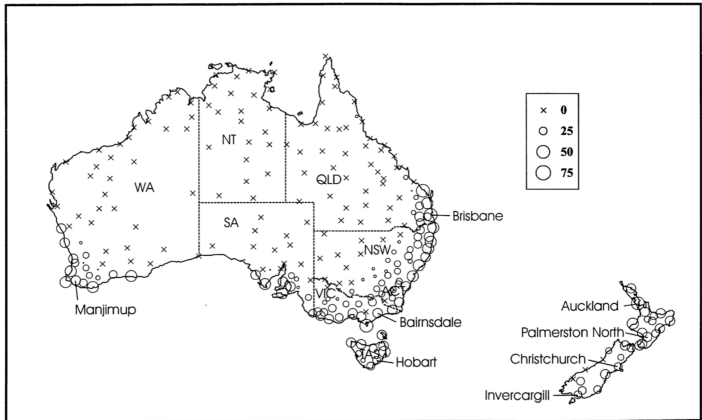
5.3.2 ASIAN GYPSY MOTH (LYMANTRIA DISPAR)

Gypsy moths (*L. dispar*) are voracious leaf feeders. They include several subspecies whose range together covers Europe, Africa, Asia, North America and South America. The Asian gypsy moth (AGM, including *Lymantria dispar asiatica*, *Lymantria dispar japonica*, *Lymantria albescens*, *Lymantria umbrosa*, and *Lymantria postalba*) is a particularly noteworthy biosecurity threat because – in contrast to European gypsy moths – the females are active fliers, capable of flying up to 30km, and therefore spreading throughout an invaded range (Matsuki *et al.* 2001; USDA 2015). It has a very broad host range. AGMs are genetically distinct from European gypsy moths so are separable with DNA tests (Carroll and Marks, 2012), though hybrids of the two can form. The native range of the AGMs include Russia, China and Japan (Carroll and Marks 2003). It also occurs in Europe along with the European Gypsy Moth, and has been found in the USA, though it is not established outside of its native range (Matsuki *et al.* 2001).

Many models of gypsy moths focus on the European moths (Allen *et al.* 1993; Gevrey and Worner 2006; Pitt *et al.* 2007; Régnière *et al.* 2009). Those specifically targeting Asian moths include those of Peterson *et al.* (2007) and Matsuki *et al.* (2001) tabulated below (Table 5)

Table 5. Gypsy moth (*Lymantria dispar*) models in the literature.

Source	Model details	Mapped predictions
Peterson <i>et al.</i> (2007)	GARP. Used 43 voucher specimens from East Asia only (they excluded data from western and central Asia because couldn't assess flight capability). Predictors: elevation, slope, aspect, annual precipitation, annual temp, mean max monthly temp, mean min monthly temp, solar radiation. 0.1 degree. Compound topographic index also used in most models.	 <p>Figure 3 Worldwide projection of native-range ecological niche model of Asian <i>Lymantria dispar</i>, showing areas globally that fit the ecological niche profile of the species as characterized on its native range. Note that Australia is included based on a separate suite of models because the compound topographic index coverage is not available for that continent. Darkest grey represents areas with complete agreement of 10 'best subsets' of GARP models, lighter grey shows lower agreement, white areas were not predicted by our model.</p>

Source	Model details	Mapped predictions
Matsuki <i>et al.</i> (2001)	<p>CLIMEX. If information from the literature didn't discriminate between AGM and EGM they used all records; if it did, just used the AGM info. We interpret the methods as: they fitted the model so it fitted both European and Asian moths.</p> <p>See predictions to right. The authors comment:</p> <p>"agreement between the observed and predicted distribution shows the poorest match in south-east China, where there is a subtropical climate ... It was not possible to assemble a set of CLIMEX parameters that were able to predict this area and at the same time predict the major part of the gypsy moth distribution. Heat stress is an important excluding factor... The high temperatures restrict the growth during the period when the insect is not in diapause.</p> <p>The model also predicts the presence of AGM throughout dryer inland parts of China and southern Russia where there are scarce records. [In Morocco] diapause is completed by February and growth is restricted by moisture after May...The overall model for AGM corresponds best at the northern cold limits and the southern dry limits. It is weakest for subtropical areas."</p>	 <p>Figure 2 CLIMEX model predictions of the distribution of Asian gypsy moth, <i>Lymantria dispar</i> (circles). The recorded distribution is shown as the shaded area (modified from Giese & Schneider (1979)). Crosses indicate climate stations where Asian gypsy moth is predicted not to survive. The size of circle indicates the degree of suitability of the climate.</p>  <p>Figure 4 Predicted distribution of Asian gypsy moth in North America. The shaded areas show the area where European gypsy moth has been recorded. Crosses indicate climate stations where Asian gypsy moth is predicted not to survive. The size of circle indicates the degree of suitability of the climate.</p>  <p>Figure 5 Prediction of the distribution of Asian gypsy moth in Australia and New Zealand. Crosses indicate climate stations where Asian gypsy moth is predicted not to survive. The size of circle indicates the degree of suitability of the climate.</p>

Asian gypsy moth: bio01 + bio12

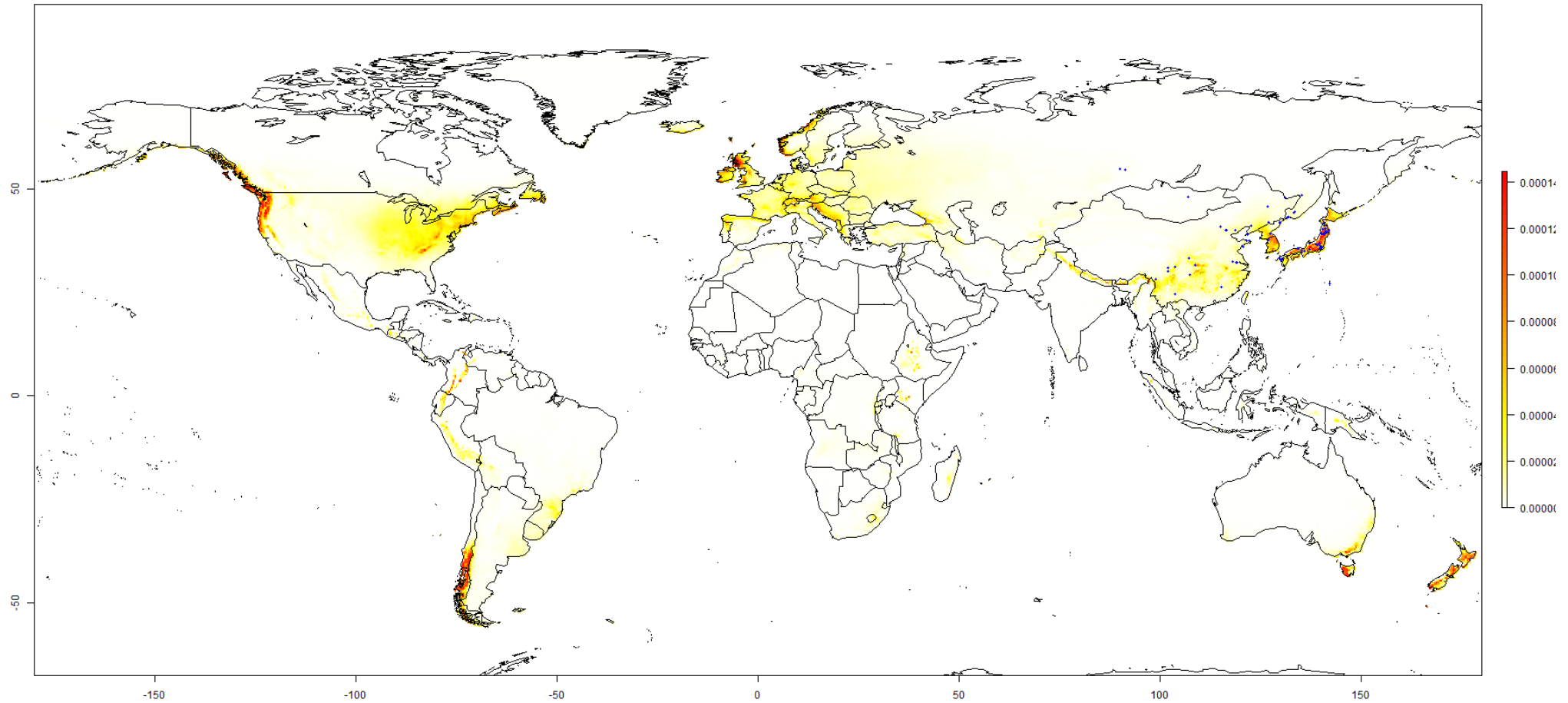


Figure 12. Best fitting GAM to BIOCLIM temperature (BIO1) and precipitation (BIO12) for Asian gypsy moth (*Lymantra dispar*). Blue crosses are recorded locations used in model development with a continental scale background.

The results for the best fitting BIOCLIM temperature (BIO1) and precipitation (BIO12) model for the distribution of Asian Gypsy Moth are shown in Figure 12. The projection to Australia predicts high suitability in the coastal regions of southern New South Wales, Victoria and much of Tasmania (Figure 13). Across the Tasman Sea, virtually all of the mountainous region of New Zealand is considered of moderate to high relative suitability (Figure 13).

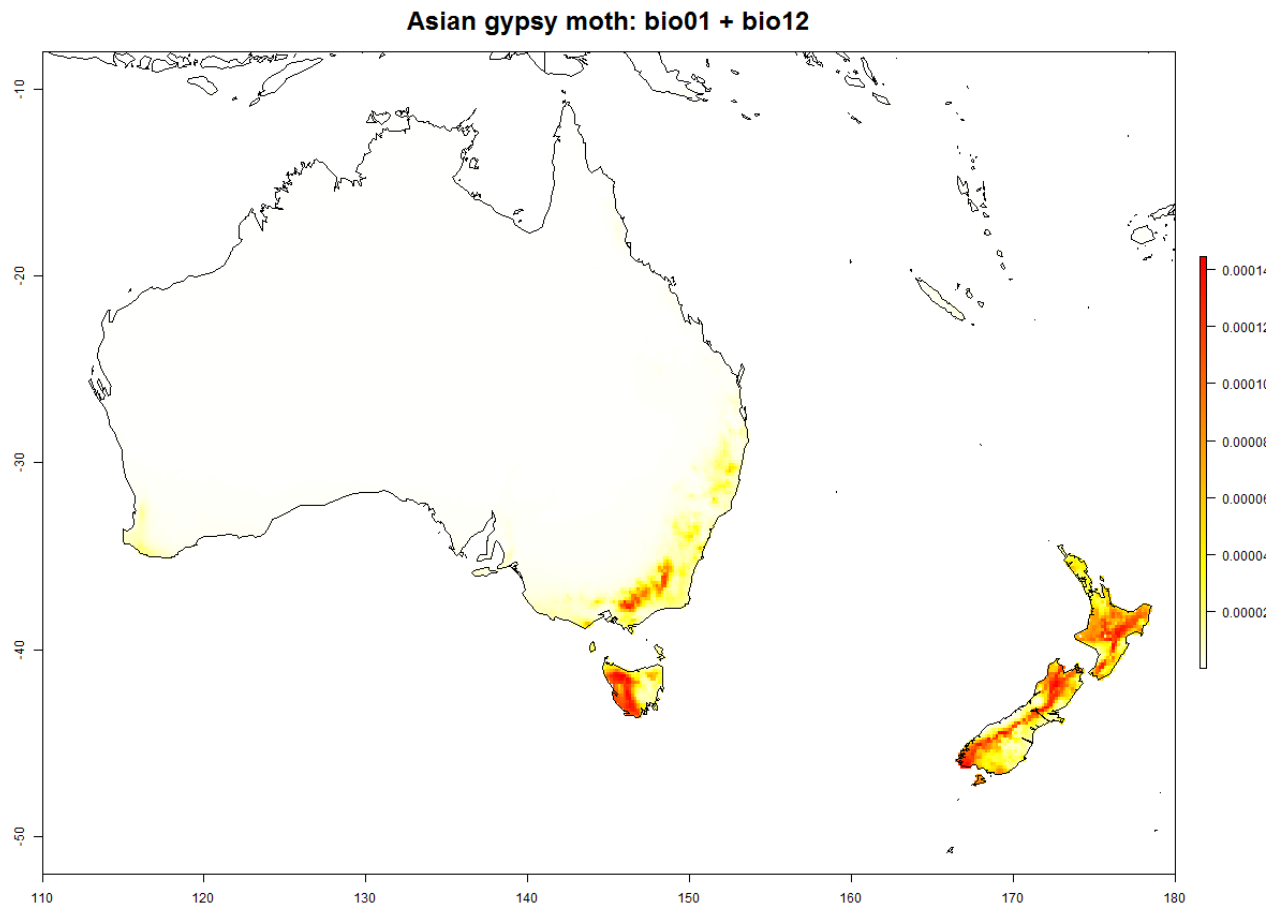


Figure 13. Projection of best fitting GAM using BIOCLIM temperature (BIO1) and precipitation (BIO12) for Asian Gypsy Moth to Australia and New Zealand. Continental background.

If we take those same variables, and project using an alpha hull approach, the projection changes again (Figure 14). Notably, areas in western Tasmania and the south-west of the New Zealand change from “hot” to “not”. A probable explanation for this is the unbounded nature of the GAM creating edge type effects arising from the extrapolation in the GAM projection. Both these areas have high rainfall, possibly outside that observed in the native range, and hence not included in the hull.

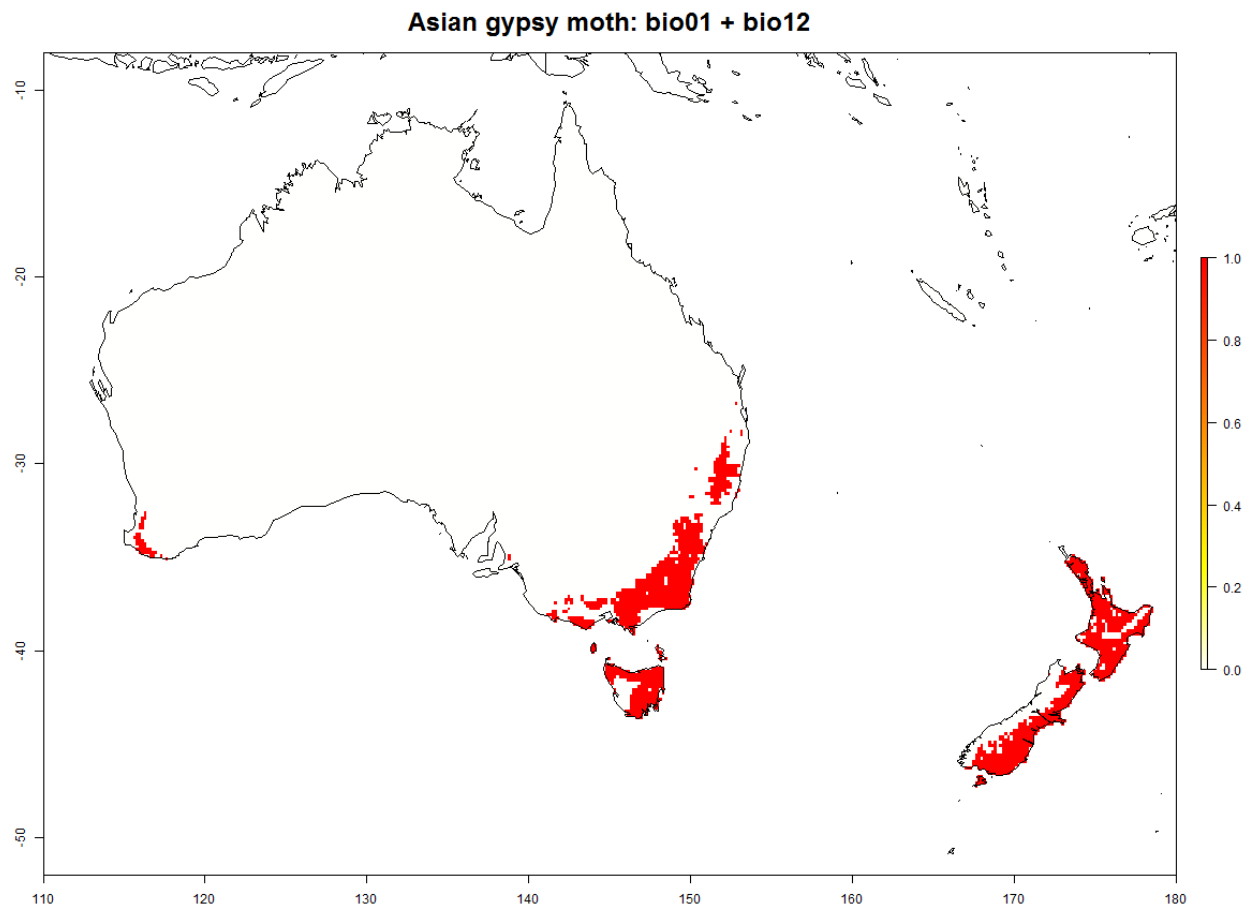


Figure 14. Project alpha-hull model using of best fitting GAM parameters BIOCLIM 01 (Annual Mean Temperature) and BIOCLIM 12 (Annual Precipitation) for Asian Gypsy Moth. Continental background.

If we use expert chosen variables, however, we get a different projection, both in Australia and New Zealand (Figure 15). In general terms, the risk map for Australia is similar, however for New Zealand it is different, with the West Coast of the South Island considered the most favourable by a large margin in the continuous predictions (Figure 15). Note though that those west-coast predictions must be extrapolations because they are masked out when hulls are used to bound the predictions (Appendix 4, Figure 2).

Finally, if we generate a best fitting (including appropriate penalization) GAM from the first 19 BIOCLIM variables we get a model whose projection contains very low suitability for both Australia and New Zealand (Figure 16).

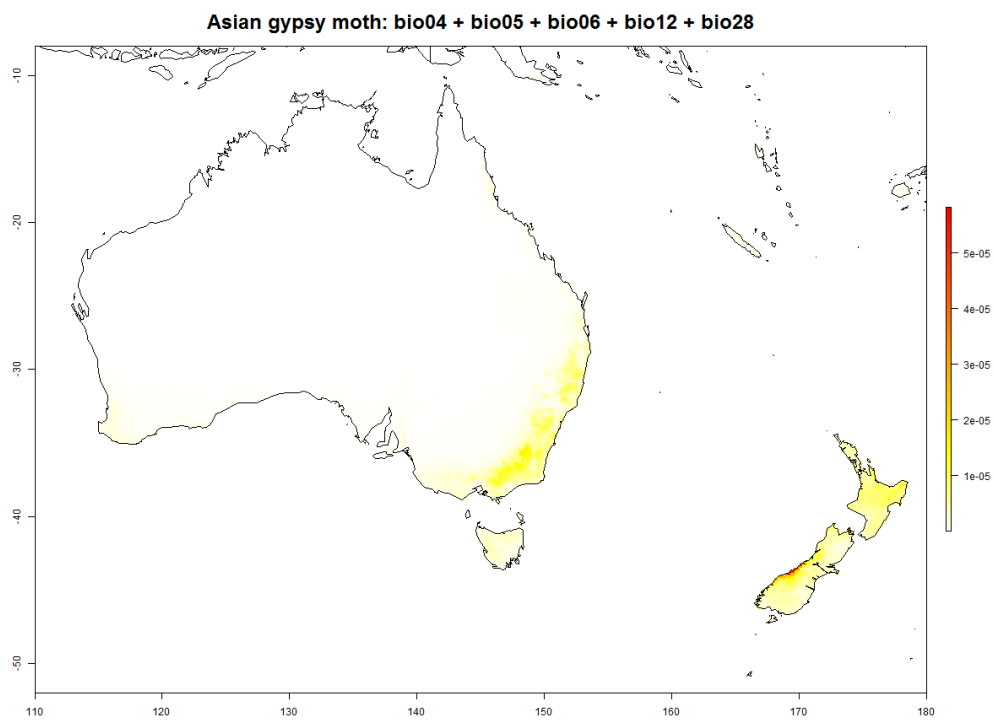


Figure 15. Projection of expert based GAM using BIO4 (Temperature Seasonality), BIO5 (Maximum Temperature of Warmest Week), BIO6 (Minimum Temperature of Warmest Week), BIO12 (Annual Precipitation), and BIO28 (Annual Mean Moisture Index) for Asian Gypsy Moth to Australia and New Zealand. Continental background.

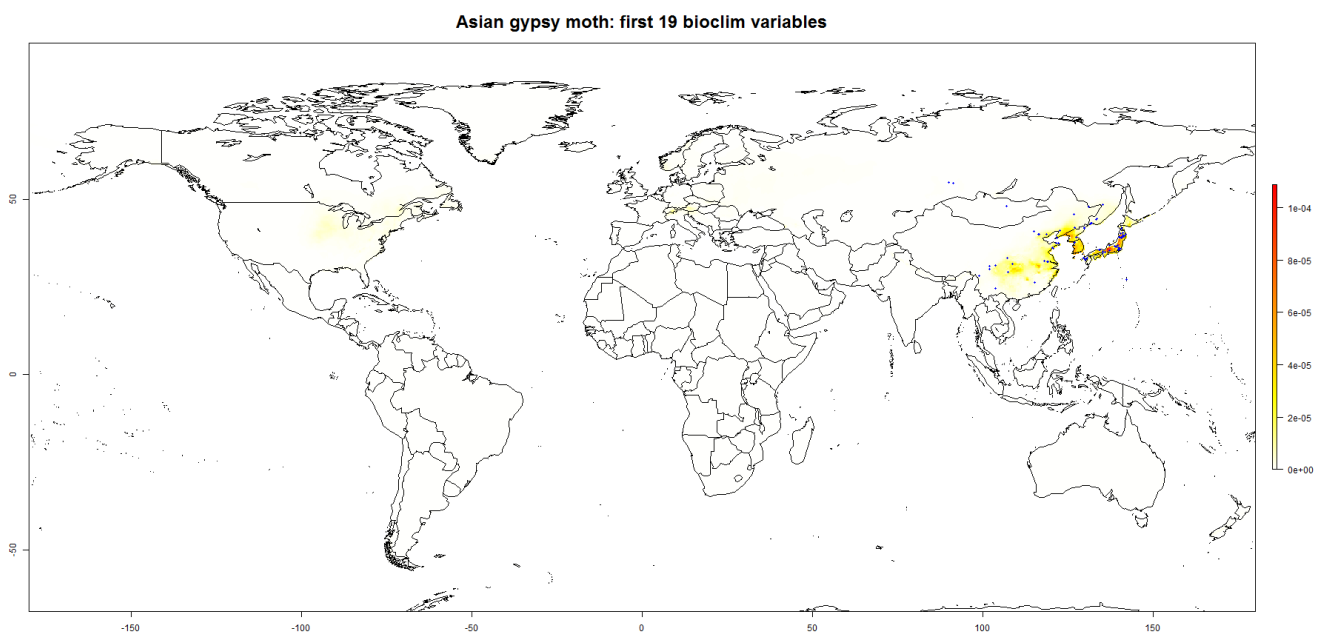


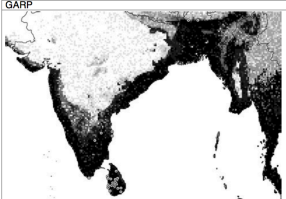
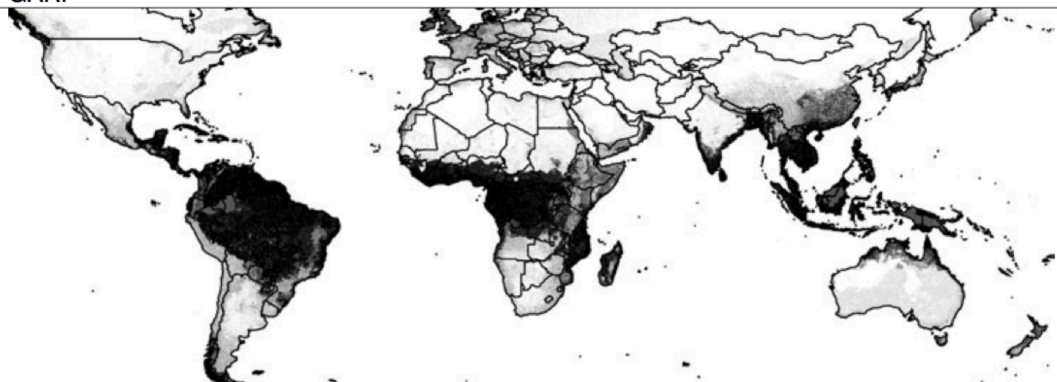
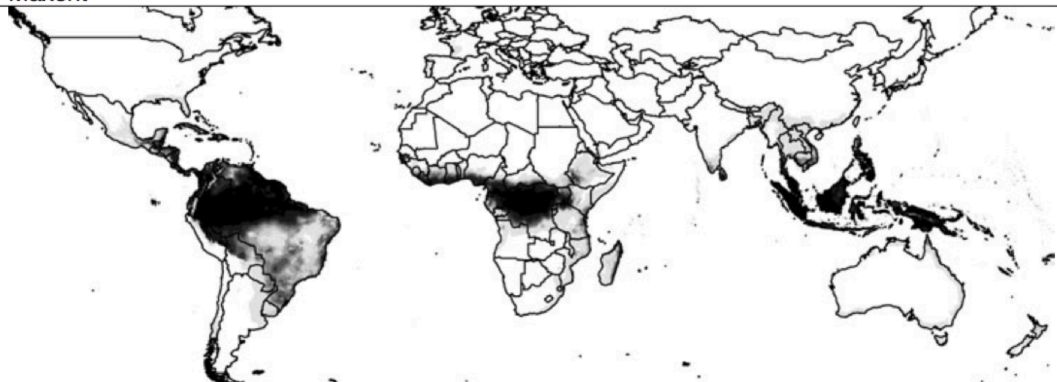
Figure 16. Best GAM derived from first 19 BIOCLIM variables. Continental background.

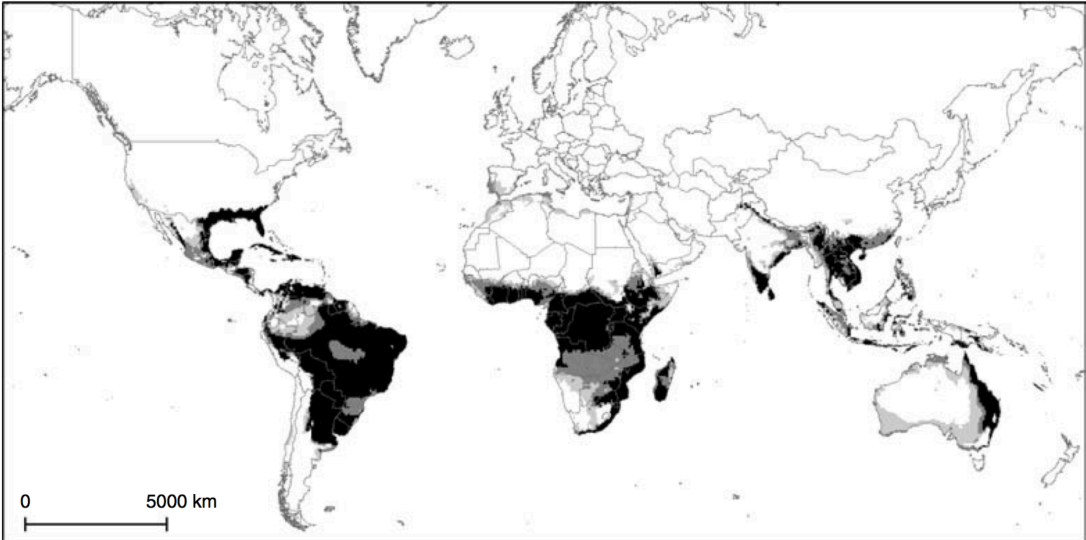
5.3.3 THE INVASIVE / ORIENTAL FRUIT FLY (*BACTROCERA INVADENS*, *B. DORSALIS*, *B. PAPAYAE*, AND *B. PHILIPPINENSIS*)

Following Schutze *et al.* (2015) we will here treat *B. invadens* and *B. dorsalis* as one species complex (also subsuming *B. papayae* and *B. philippinensis*). For simplicity we refer to it as the Oriental fruit fly. Oriental fruit fly is a damaging pest of fruit and vegetable species because of its wide host range and ability to attack some fruit green. A serious pest worldwide, the Queensland Government (2015) reports: “Oriental fruit fly is endemic in southeast and southern Asia and has spread to Hawaii, Tahiti, Mariana Islands and Africa. It has been present in Papua New Guinea since 1992. In March 1993, it was detected for the first time in Australian territory on the islands of Saibai, Boigu and Dauan, adjacent to the Papua New Guinea coast; and on Stephen and Darnley Islands close to the centre of Torres Strait and was subsequently eradicated”. An incursion on mainland Australia (identified then as *B. papayae*) occurred in 1995 near Cairns, with eradication declared in 1999 ((Cantrell *et al.* 2002)).

Previous models for either species include those of De Meyer *et al.* (2010) and Stephens *et al.* (2007) tabulated below (Table 6) (see species records used, below table). Key results to note are the major differences in the projected suitability in Australia depending on the modelling approach and data used. Note also that the distribution records used reflect previous concepts of two separate species: De Meyer *et al.* (2010) (Figure 17) and Stephens *et al.* (2007) (Figure 18) (right, native = circle, x = invasive). The model of Stephens *et al.* (2007) does a pretty good job of mimicking the distribution *B. invadens* (now considered the same species) over some parts of Africa. It does, however, appear to over-predict into southern parts of Africa based on known presence data (assuming that the species has had time to disperse to all suitable locations), which raises the question as to whether the prediction into south-eastern Australia is sound.

Table 6. Oriental fruit fly (*Bactrocera dorsalis* complex) models in the literature.

Source	Model details	Mapped predictions
De Meyer <i>et al.</i> (2010)	<p>Models only fitted to native range data for <i>B. invadens</i>, identified from vouchered specimens (34 records across India, Sri Lanka and Bhutan, mostly in Sri Lanka – See Figure 17). Predictors: WorldClim variables at 1km resolution: annual mean temperature, mean diurnal range, maximum temperature of warmest month, minimum temperature of coldest month, annual precipitation and precipitation of the wettest and driest months (=variables 1, 2, 5, 6, 12, 13, 14). Used Maxent and GARP (both with defaults) – not clear what background used, but possibly the extent shown here:</p> 	<p>GARP</p>  <p>Maxent</p> 

Source	Model details	Mapped predictions
Stephens <i>et al.</i> (2007)	CLIMEX. Based on <i>B. dorsalis</i> ; records shown in Figure 18.	 <p data-bbox="1016 772 2130 820">— The climate suitability (EI) for the oriental fruit fly under the reference climate (1961–1990 averages) projected using CLIMEX™ (□, unsuitable (0.00–0.49); ■, marginal (0.50–9.99); ■, suitable (10.00–19.99); ■, optimal (20.00+)).</p>

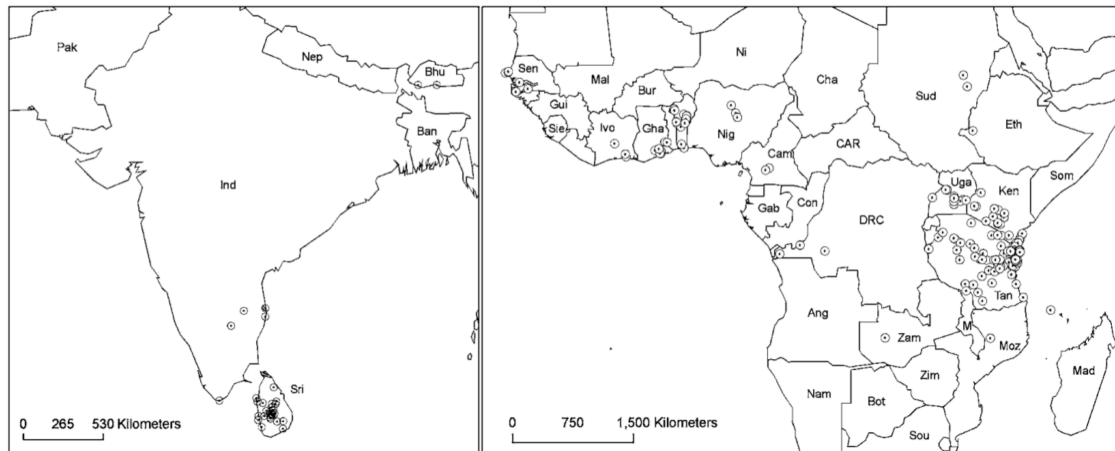


Fig. 1. Distribution records for *B. invadens*. Native records in India (Ind), Sri-Lanka (Sri) and Bhutan (Bhu). Non-native records in Africa.

Figure 17. Distribution records for *B. invadens* used by De Meyer *et al.* (2010).

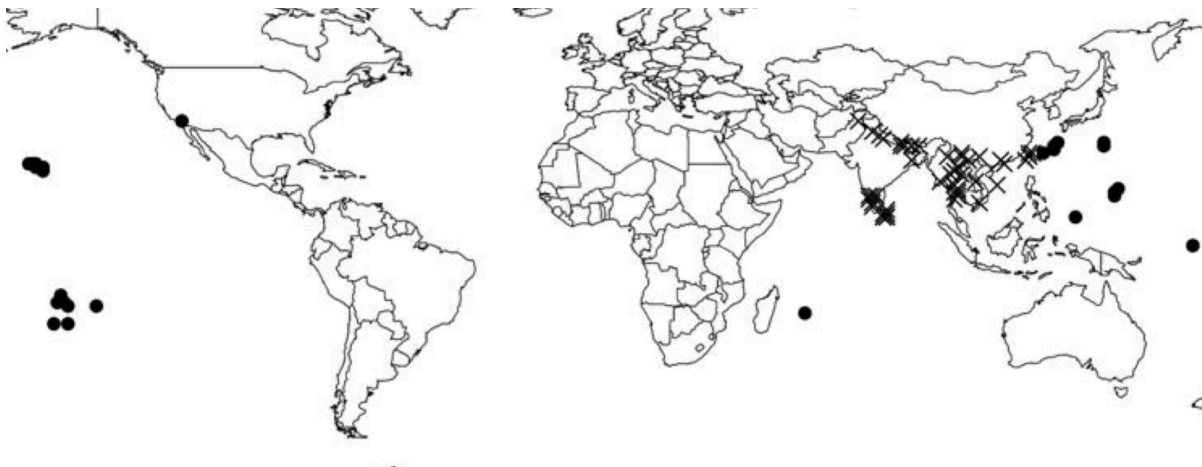


Figure 18. From Stephens *et al.* (2007), the distribution records for *Bactrocera dorsalis* used in their modelling (circles, native range; crosses; invaded) – as then considered a separate species to *B. invadens*.

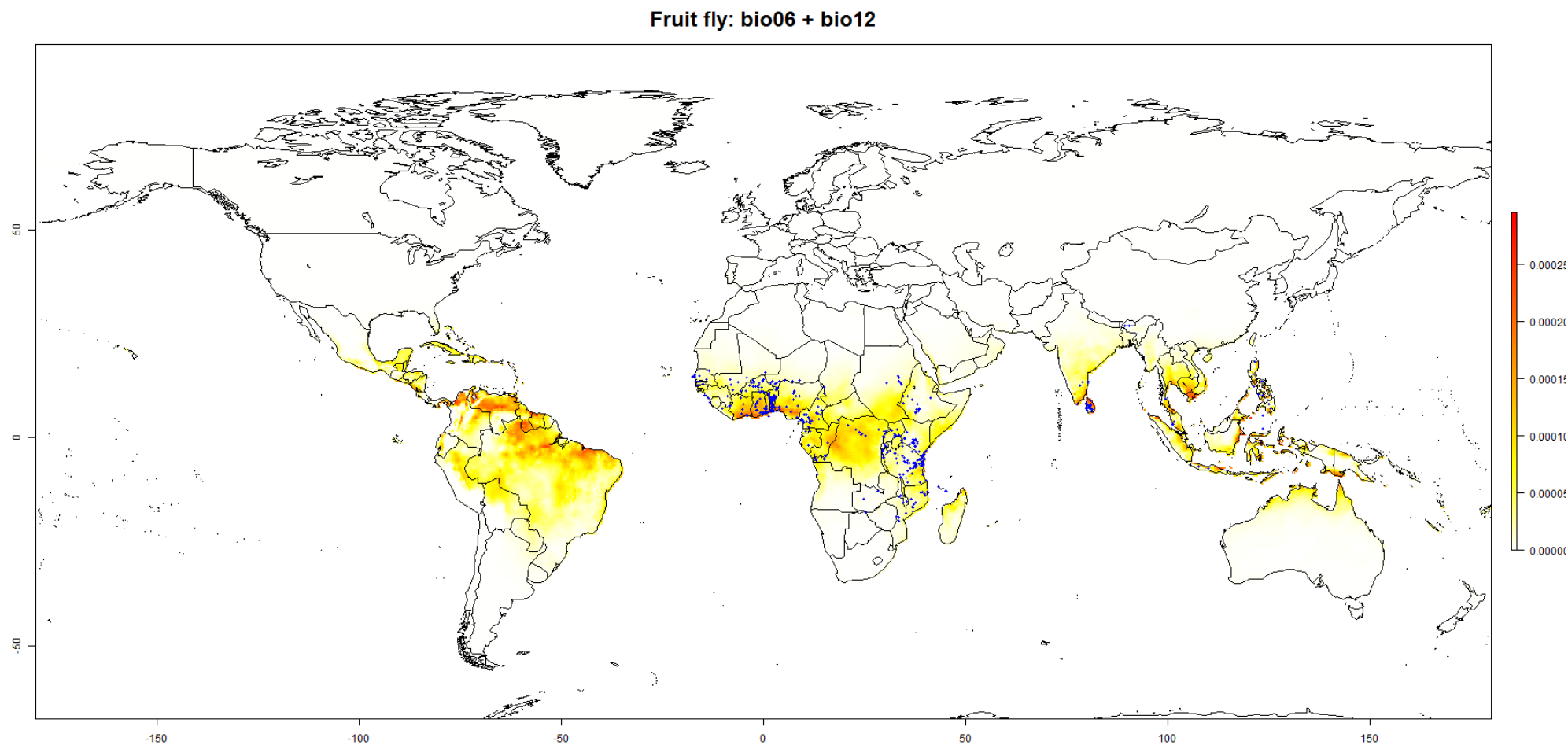


Figure 19. Best fitting GAM using BIOCLIM 06 (Min. Temp. of Coldest Period) and BIOCLIM 12 (Annual Precipitation) for the global distribution of the *Bactrocera dorsalis* complex. Blue crosses are reported collections. Background is continental.

The results for the best fitting BIOCLIM temperature and precipitation GAM for the distribution of the *Bactrocera dorsalis* complex are shown in Figure 19. Of note is the observed occurrence along the Nile River that is not well predicted, presumably on account of irrigation. The SDM is most similar to the MAXENT model of De Meyer *et al.* (2010) (Table 6)

The projected Australian distribution of the *Bactrocera dorsalis* complex predicts highest suitability north of Cairns, Kimberley Coast and the far Top End of the Northern Territory. The area of highest suitability is marginally consistent with the incursion near Cairns in 1995 (Cantrell *et al.* 2002). This differs considerably from the CLIMEX projection of Stephens *et al.* (2007) (Table 6) that predicted a widespread population through QLD into NSW (Table 6).

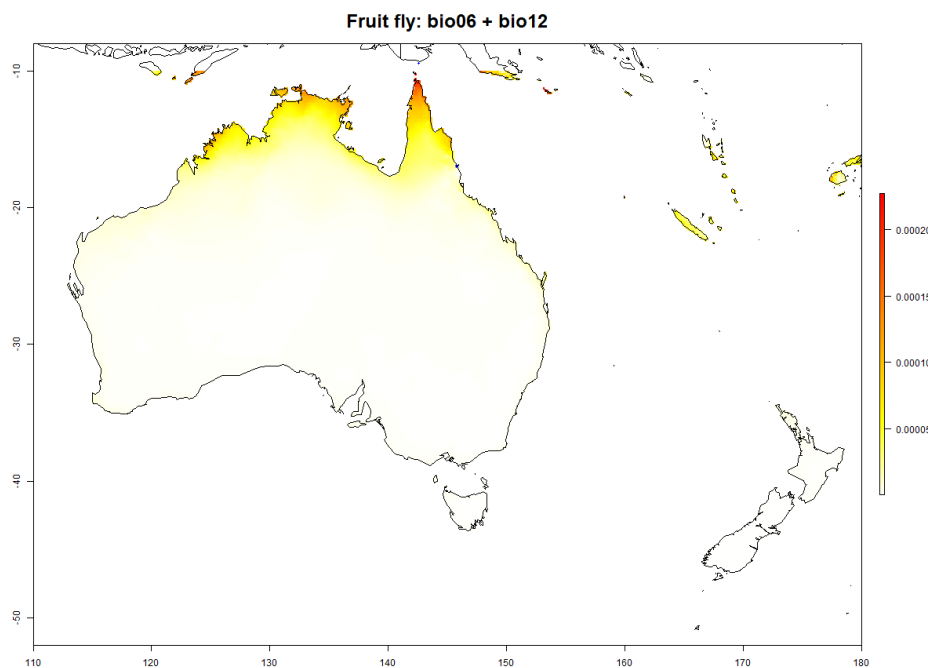


Figure 20. Projection of best fitting GAM using BIOCLIM O6 (Min. Temp. of Coldest Period) and BIOCLIM 12 (Annual Precipitation) for the *Bactrocera dorsalis* species complex to Australia and New Zealand. Continental background.

The alpha-hull model corresponding to the GAM projects a qualitatively similar result, with a potential distribution down the entire eastern seaboard of Australia and into parts of Northland, New Zealand (Figure 21). Using a bounding box c.f. an alpha hull for the same variables generates a projection that would be considered untenable, given the extensive distribution in desert regions. (Figure 22).

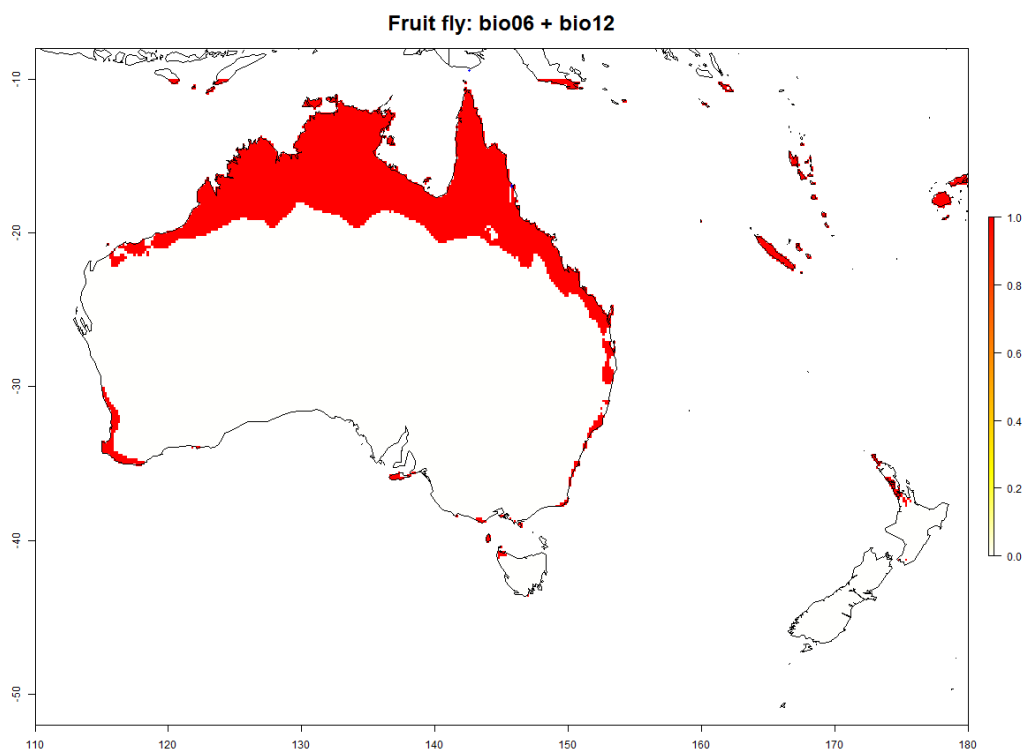


Figure 21. Alpha hull model based on best fitting GAM variables BIOCLIM O6 (Min. Temp. of Coldest Period) and BIOCLIM 12 (Annul Precipitation).

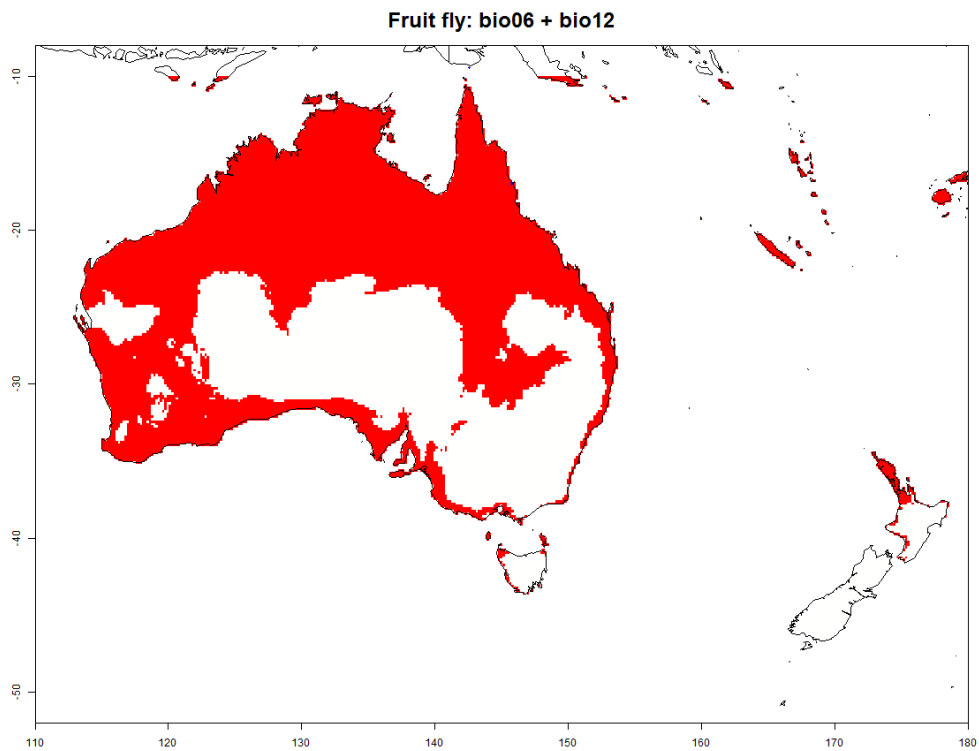


Figure 22. Bounding box model based on best fitting GAM variables BIOCLIM O6 (Min. Temp. of Coldest Period) and BIOCLIM 12 (Annul Precipitation).

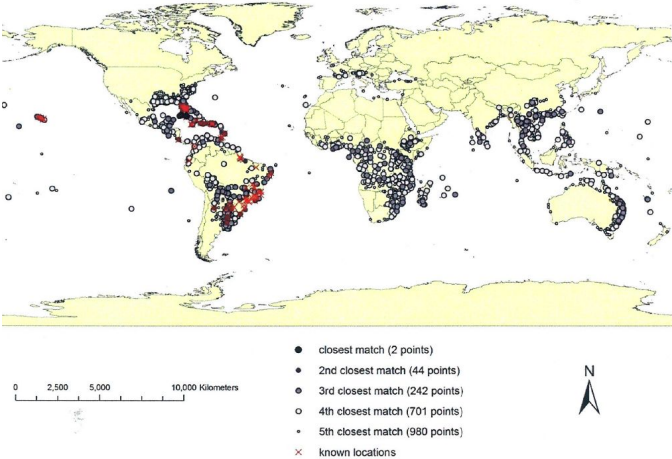
5.3.4 MYRTLE/GUAVA RUST (*PUCCINIA PSIDII* S.L.)

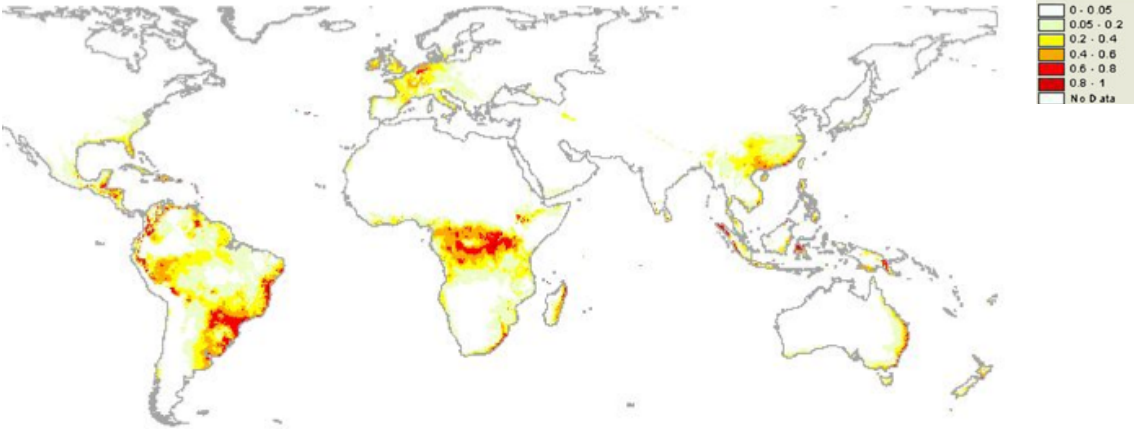
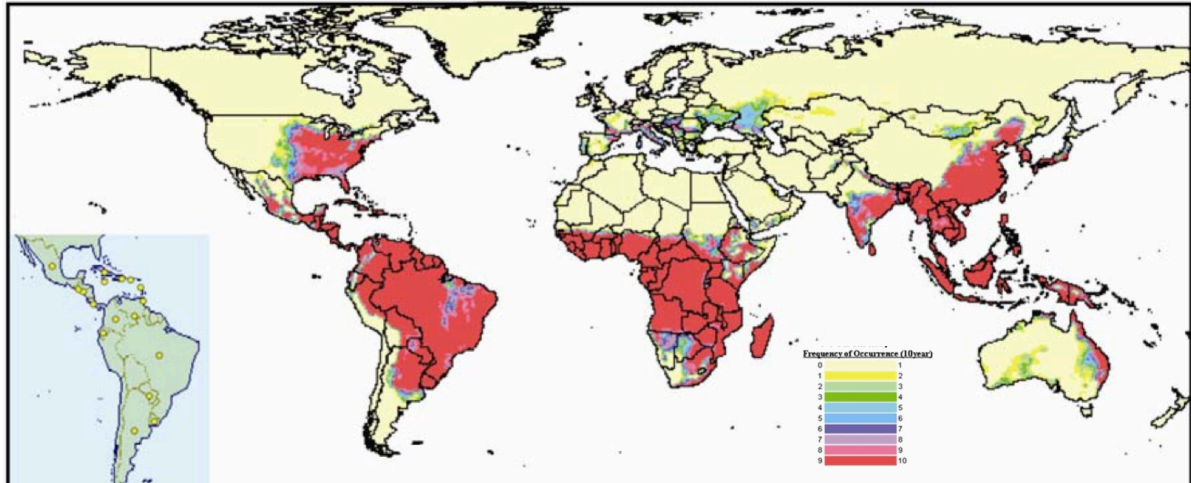
Background

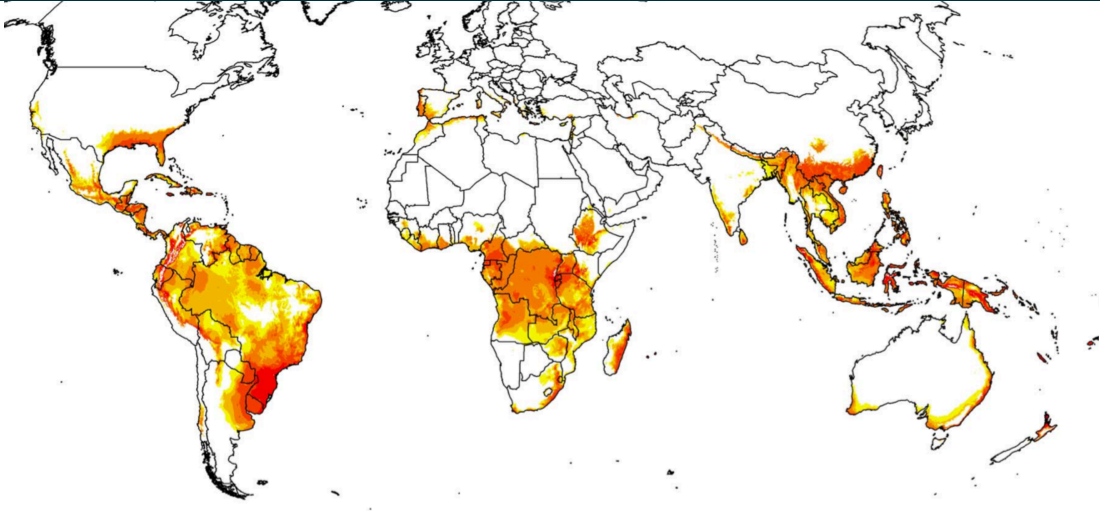
Myrtle rust or guava rust (*Puccinia psidii sensu lato*) has a large and varied host range with sometimes large impacts on infected species. Whilst there has been debate over taxonomy (Carnegie and Cooper 2011; Glen *et al.* 2007; Simpson *et al.* 2006), here we use the broad definition of the *P. psidii* complex, which includes *Uredo rangellii*. The native range of the species is South and Central America, including Brazil, Argentina, Uruguay, Jamaica and Puerto Rica (Elith *et al.* 2013, Online Appendix 1).

Several models have been developed for predicted distributions of guava/myrtle rust including those tabulated below (Table 7).

Table 7. Myrtle rust (*Puccinia psidii sensu lato*) models in the literature.

Source	Model details	Mapped predictions
Biosecurity Australia (2009)	CLIMATE, using 16 temperature and rainfall-related variables	

Source	Model details	Mapped predictions
Elith <i>et al</i> (2013)	<p>MaxEnt, using the same 7 predictors as used in our “expert” set (representing temperature, precipitation, humidity and aridity) and settings to produce relatively smooth models (linear and quadratic features). Native and invaded range records used to fit model.</p>	
Magarey <i>et al.</i> (2007)	<p>NAPPFAS – a weather-based system that uses either daily (USA) or monthly data series to predict years of suitable conditions over a selected timeframe. For guava rust, settings were: average daily max temp $\leq 33^{\circ}\text{C}$; average daily min temp $> 13^{\circ}\text{C}$; 5-25, wet days per month. If ≥ 3 months met these conditions, then the climate would be suitable for the pathogen. The model was run with 10 years (1993–2002) of weather data</p>	

Source	Model details	Mapped predictions
Kriticos <i>et al.</i> (2013)	CLIMEX – Ecoclimatic Index shown here.	 <p>The map displays the CLIMEX Ecoclimatic Index predictions for a species, showing high suitability (red/orange) concentrated in South America, Africa, and Southeast Asia. The index is shown as a heatmap overlay on a world map, with the highest values (red) indicating the most suitable areas for the species. The map shows high suitability (red/orange) in South America, particularly in the Amazon basin and surrounding regions. In Africa, high suitability is concentrated in the central and southern parts of the continent. In Southeast Asia, high suitability is concentrated in the island nations and coastal regions. The map also shows some areas of moderate suitability (yellow) in North America, Europe, and Australia.</p>

The best fitting 2-variable GAM (fitted to only native range data, with continental background) is shown in Figure 23. This SDM has some problems. At the Australian scale the predictions are somewhat questionable, with observed occurrences not well matched by suitability projections (Figure 24).

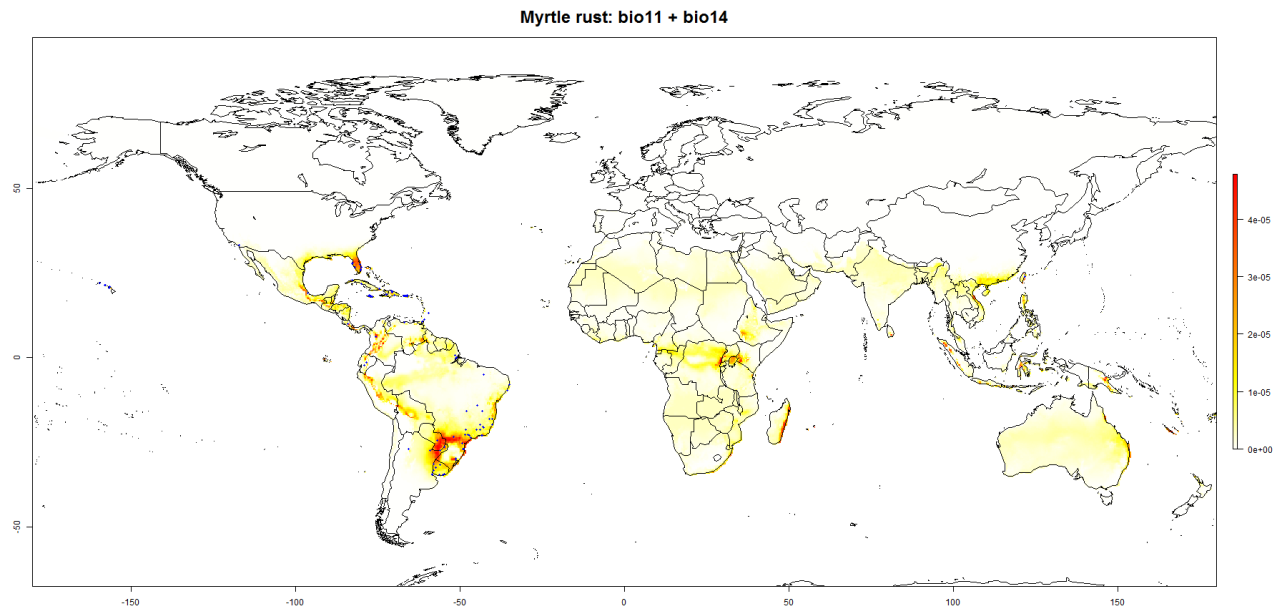


Figure 23. Best fitting GAM for myrtle rust based on BIOCLIM variables BIO11 and BIO14.

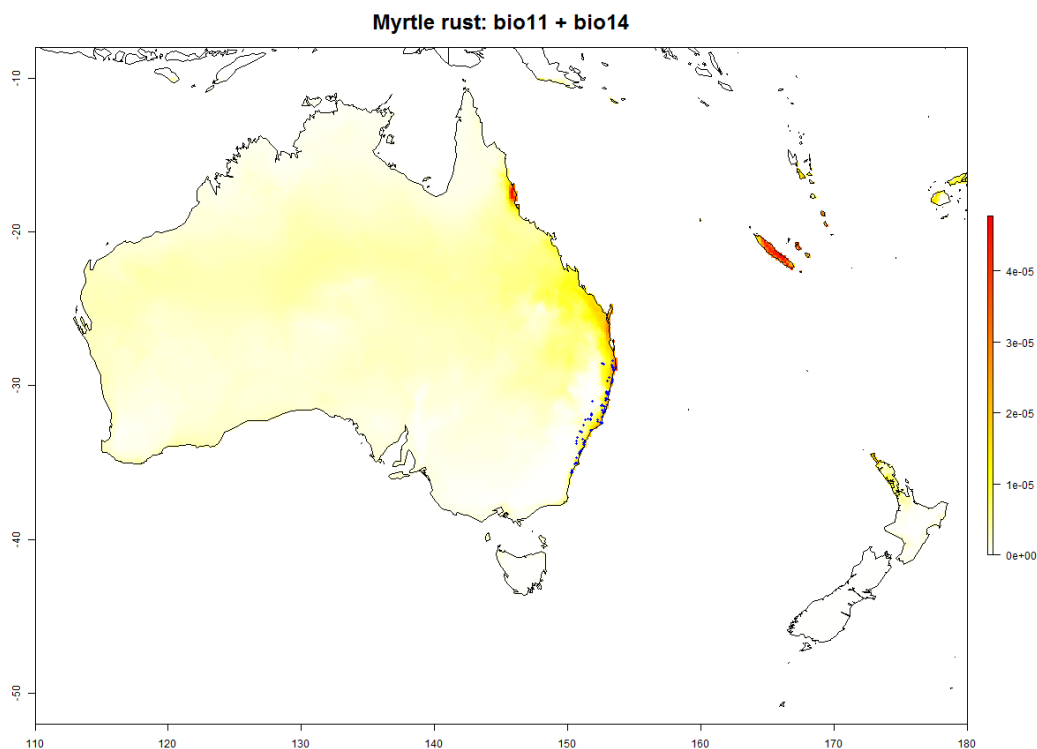


Figure 24. Best fitting GAM for myrtle rust based on BIOCLIM variables BIO11 (Mean temp. of coldest quarter) and BIO14 (Precipitation of driest period) applied to Australia and New Zealand. Observations are small blue crosses.

The alpha-hull model using the two variables chosen by the best fitting GAM, and only native range data, performs very poorly, with a high proportion of known Australian occurrences clearly omitted and large areas of central Australia that would be considered likely errors of commission (Figure 25).

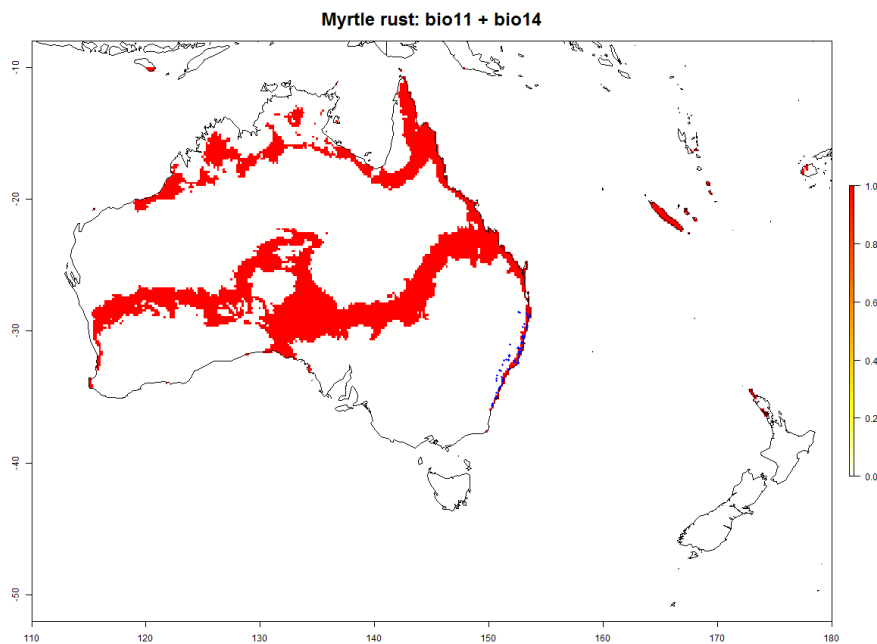


Figure 25. Alpha-bull myrtle rust model based on BIOCLIM variables BIO11 (Mean temp. of coldest quarter) and BIO14 (Precipitation of driest period). Observations are small blue crosses.

Using variables identified previously as being important doesn't seem to help, with different shortcomings (Figure 26).

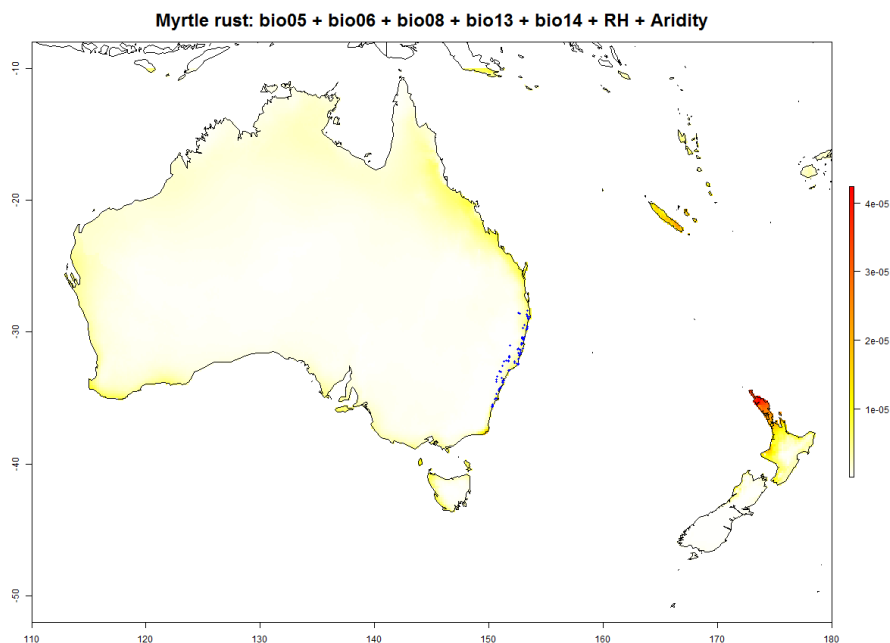


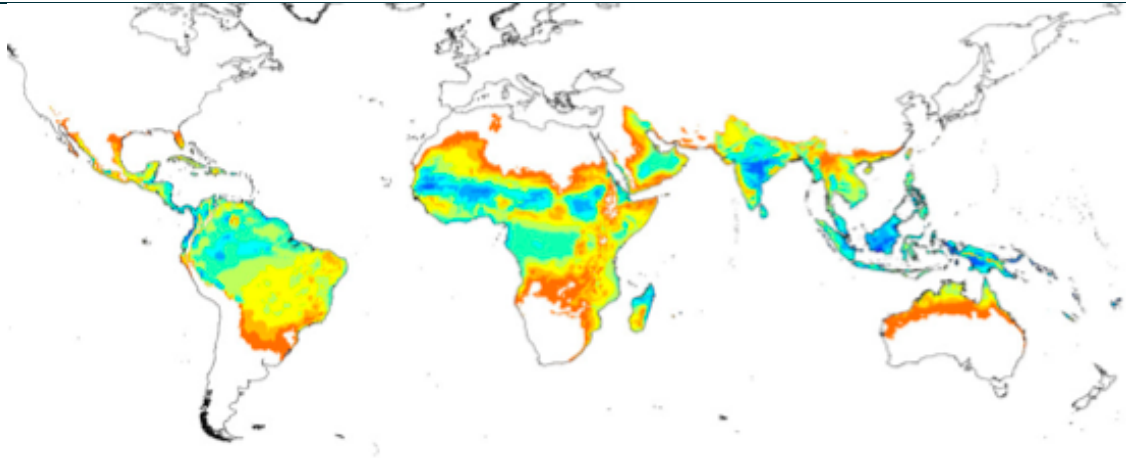
Figure 26. Projected suitability for myrtle rust from GAM based expert identified variables.

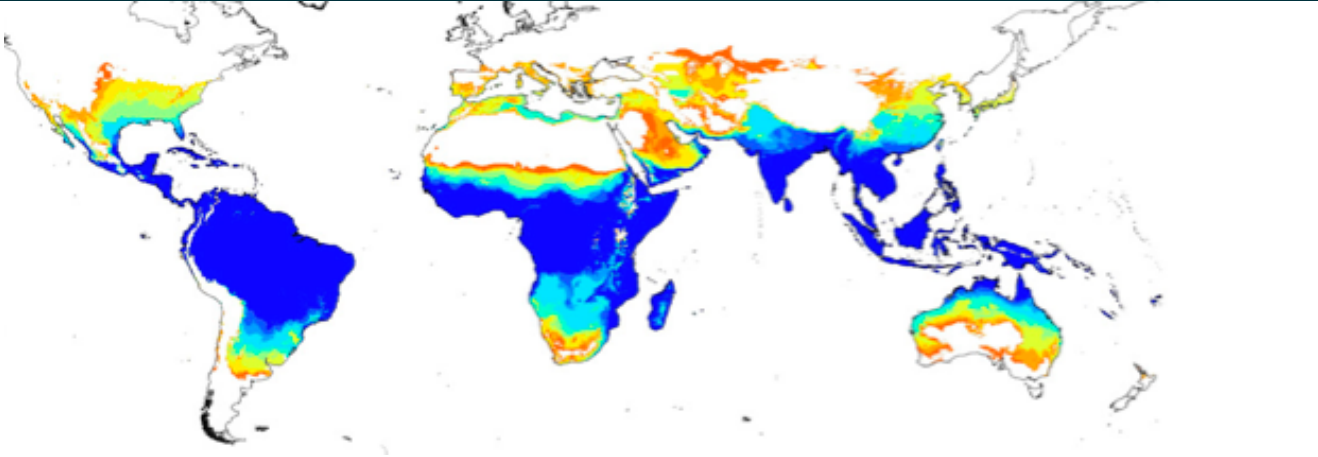
5.3.5 CANE TOADS (RHINELLA MARINA)

Background

Cane toads (*Rhinella marina*, formerly *Bufo marinus*) are a well known pest in Australia, they are a useful case study since they are relatively unaffected by biotic interactions and are well studied, with models including a mechanistic model published (Table 8). Their native range is in South America; Tingley *et al.* (2014) discuss the effect of a closely related species, *R. schneideri*, on its native distribution.

Table 8. Cane toads (*Rhinella marina*) models in the literature.

Source	Model details	Mapped predictions
Tingley <i>et al.</i> (2014)	<p>Maxent fitted to data from native-range, and 5 predictors related to heat and water balance: minimum temperature of the coldest month, maximum temperature of the warmest month, mean annual temperature, mean humidity of the warmest quarter, and precipitation of the warmest quarter.</p> <p>Predictions are depicted in 10% suitability classes ranging from white to orange to yellow to green to blue.</p>	

Source	Model details	Mapped predictions
Kearney <i>et al.</i> (2008)	<p>Mechanistic model – not fitted to any observed data; based on known physiology of the species.</p> <p>Predictions are depicted in 10 equal interval classes, with the highest class (royal blue) depicting 9– 12 breeding months per year and the white area representing no breeding months per year.</p>	 <p>The map displays global predictions for breeding months. The highest breeding frequency (9-12 months, royal blue) is concentrated in the Amazon basin of South America, central Africa, and parts of Southeast Asia and Australia. Intermediate frequencies (yellow and orange) are found in North America, Europe, and northern Africa. Large white areas, indicating no breeding, are present in high-latitude regions like Siberia and northern Canada, as well as arid zones in North and West Africa.</p>

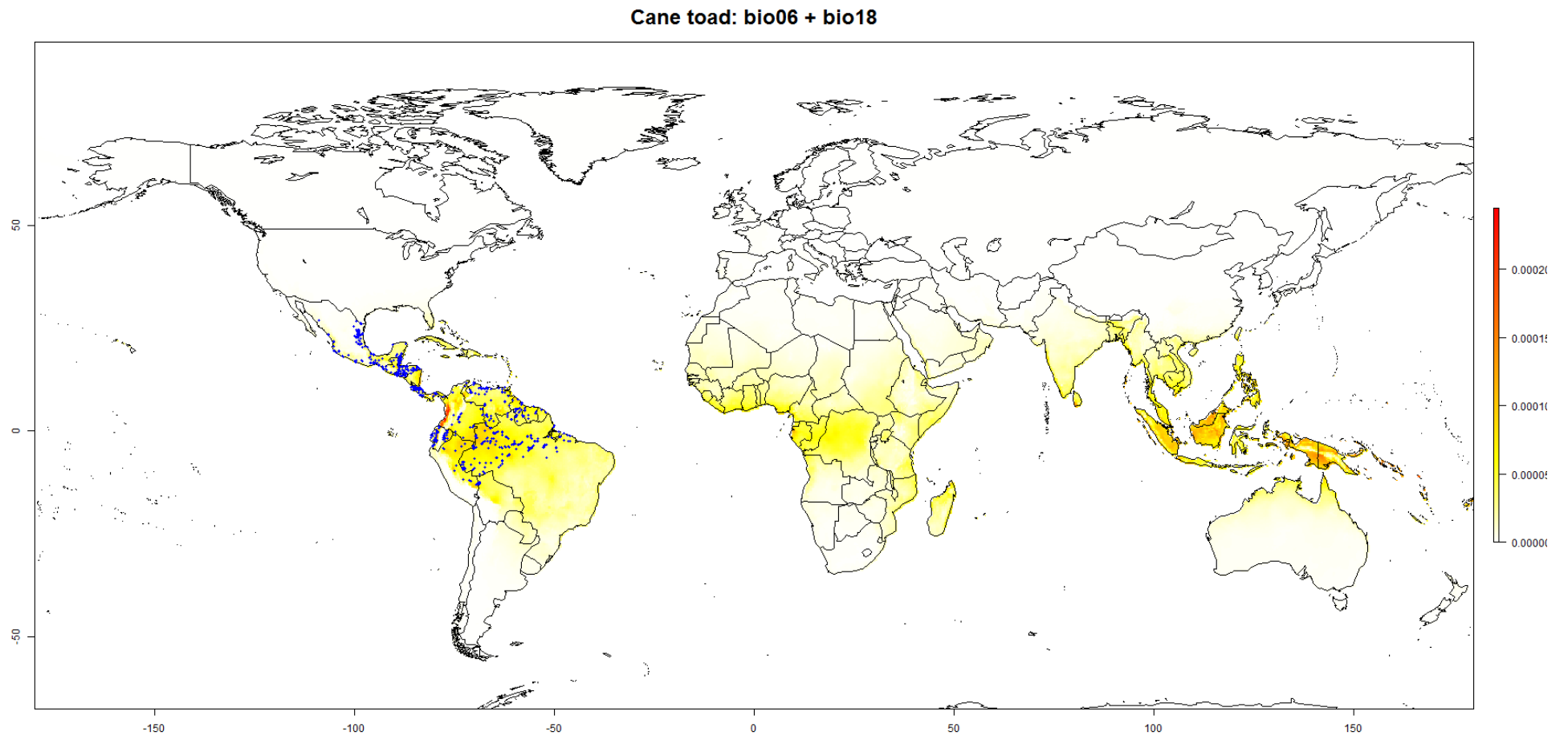


Figure 27. Best fitting GAM for the cane toad (*Rhinella marina*) based on BIOCLIM variables BIO6 (Min. Temp. of Coldest Period) and BIO18 (Precipitation of Warmest Quarter). Observations are blue crosses. Background is continental.

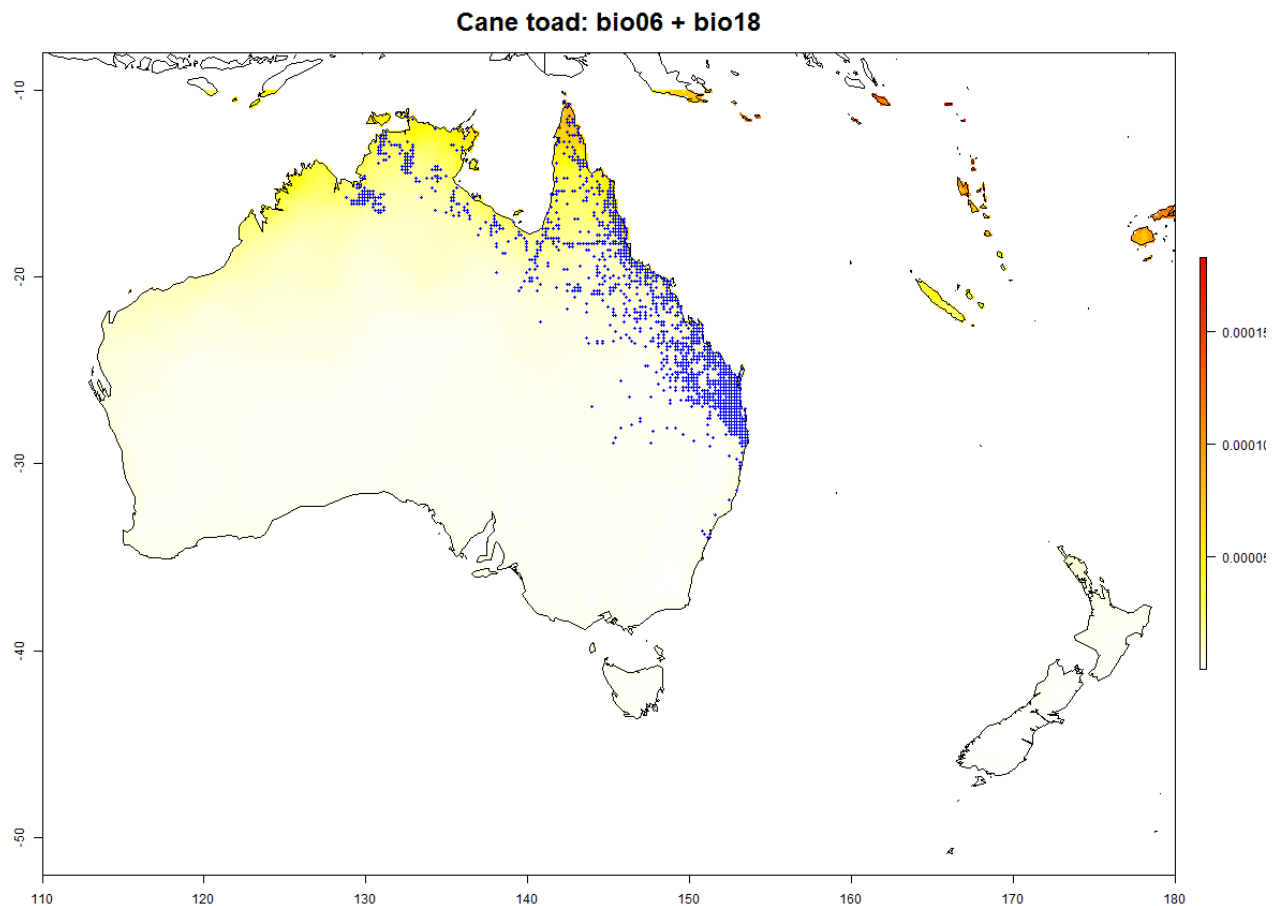


Figure 28. Projection of best fitting GAM for the cane toad (*Rhinella marina*) based on BIOCLIM variables BIO6 and BIO18 and continental background. Model is fitted to native range only.

The best fitting cane toad GAM performs particularly poorly when projected to Australia (Figure 28). While the “expert” set of predictors in combination with alpha hulls seemed to perform reasonably well for fire ants, this is not the case for cane toads, particularly in the southern parts of the invaded range (Figure 29 and Figure 30)**Error! Reference source not found..** Moving from an alpha hull type model to a bounding box approach (both fitted to native range data) generates considerable additional commission errors (Figure 31). The direction of errors, however, is not necessarily consistent, and using a bounding box approach on expert identified BIOCLIM variables results in considerable omission errors (Figure 32).

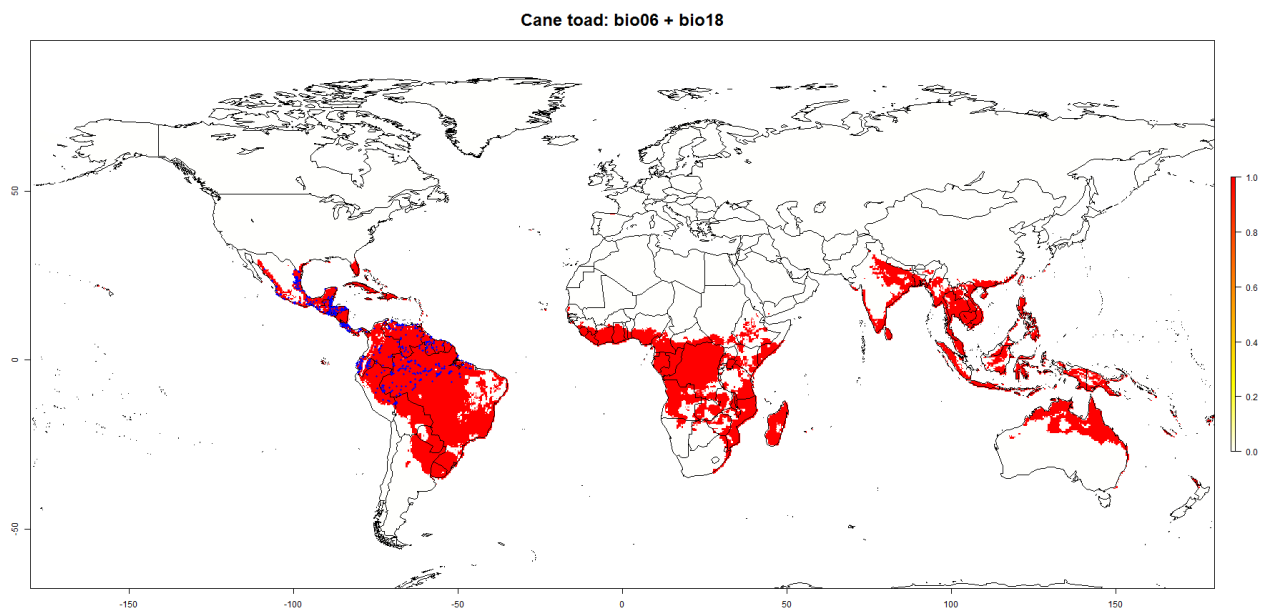


Figure 29. Alpha hull projection for cane toad, BIO06 and BIO18 (best fit GAM), continental background. Note the area of predicted suitability in South America that is inhabited by the congener *Rhinella schneideri*.

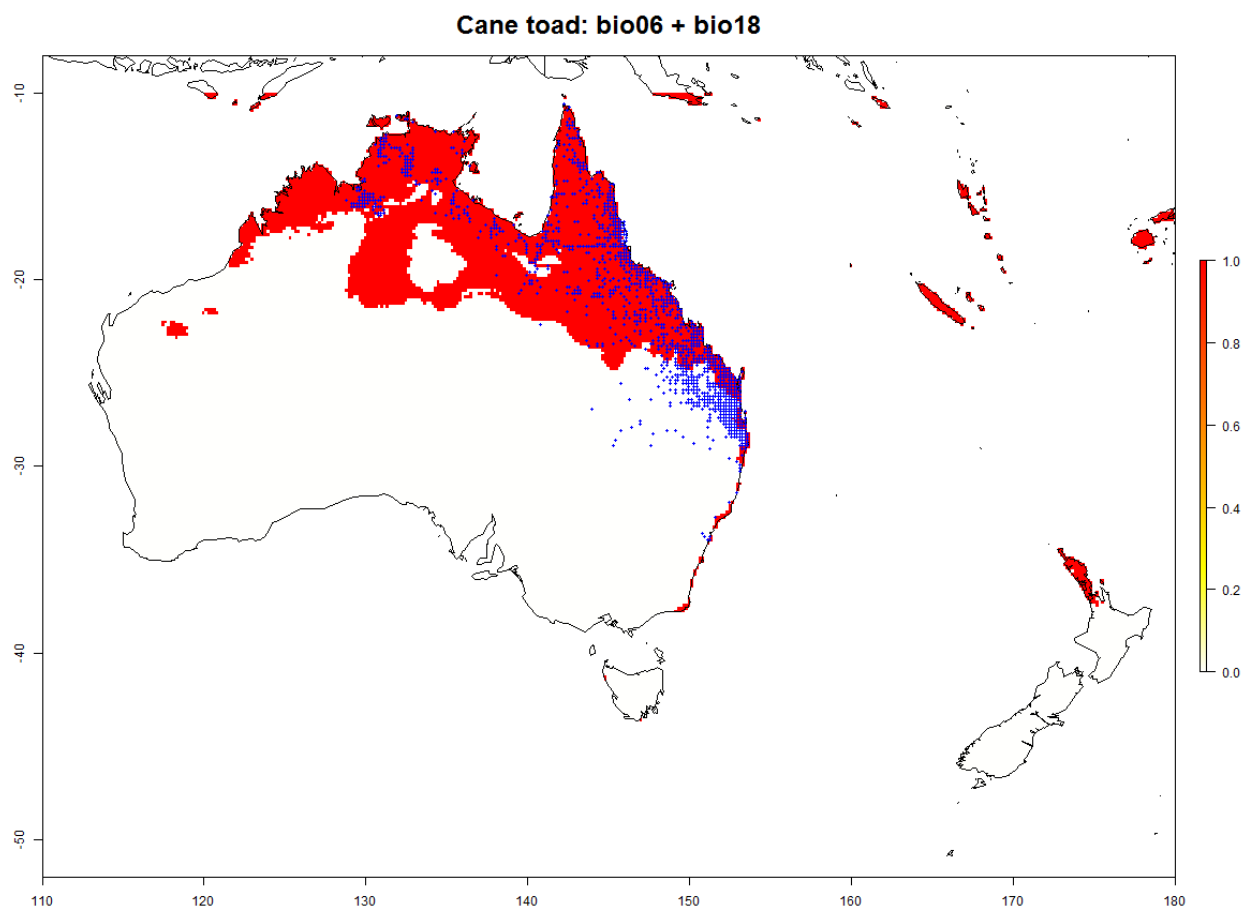


Figure 30. Alpha hull projection for cane toad, BIO06 and BIO18 (best fit GAM), continental background.

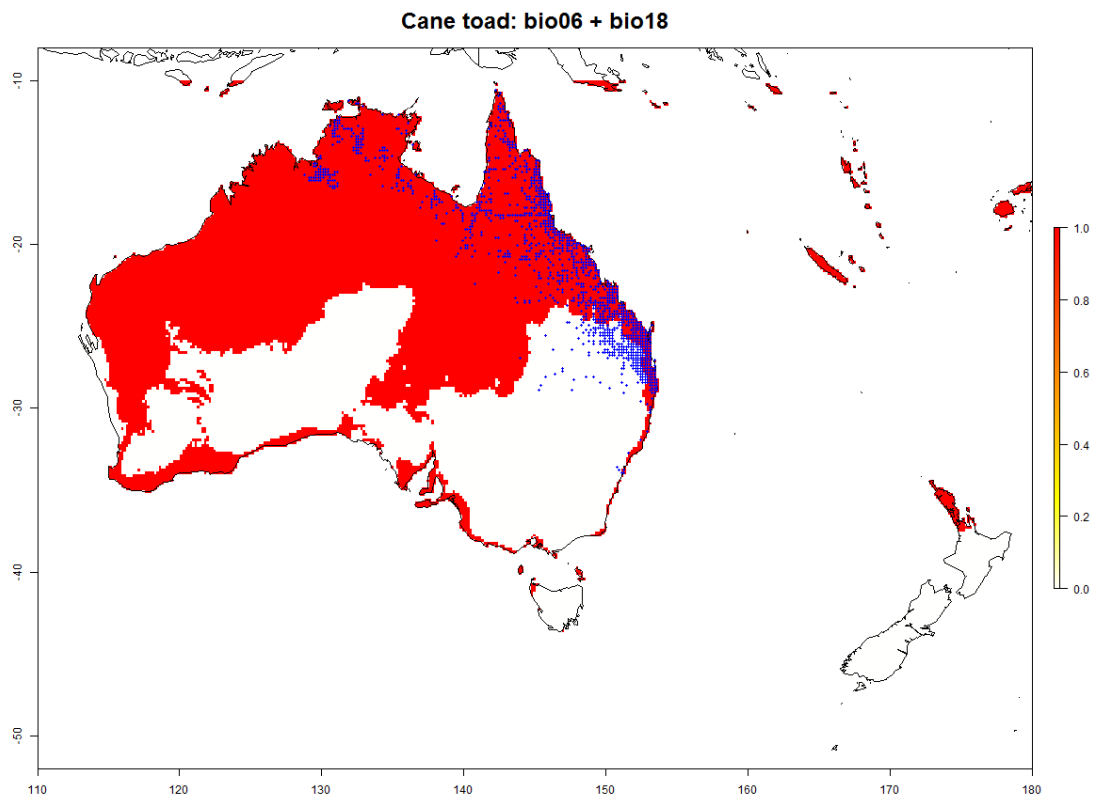


Figure 31. Bounding box projection for the cane toad, BIO06 and BIO18, continental background.

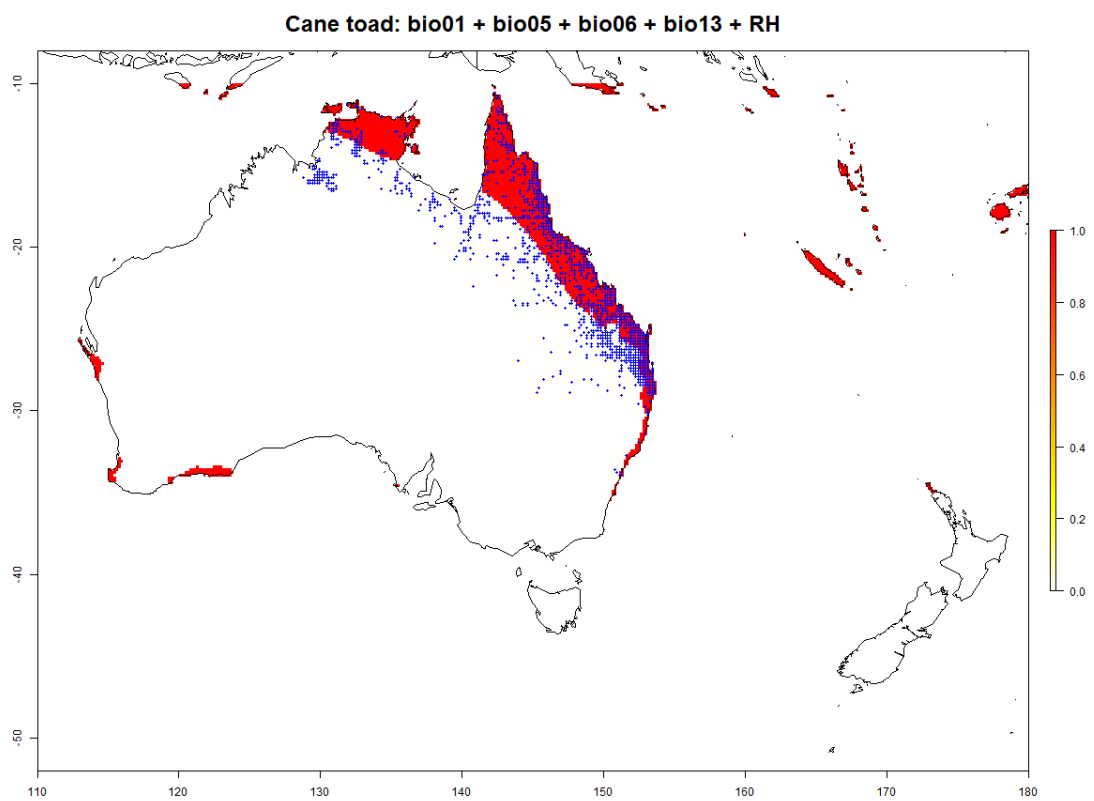


Figure 32. Bounding box projection for the cane toad, expert derived BIOCLIM variables, continental background.

6 Problems with projections

Introduction

The previous chapters have identified a lack of explicit and unambiguous information about proximal variables in the scientific literature. Based on this finding, a draft protocol was proposed to attempt to empirically identify proximal variables. This was based on trying to find variables that were strongly predictive, as the claim of proximal variables is that they will be strongly predictive regardless of geographical location.

Analysis of this protocol across the case studies identified that no one approach made consistent, reliable predictions. While some analysis choices may be worse than others there is no general automated approach that can be recommended in all circumstances.

A partial explanation for this result is found in the scientific literature. There is extensive discussion of the role of biotic interactions in confounding environmental patterns. For instance a predator or competitor may impact the distribution of a species in its native range. In the introduced range the predator or competitor may not exist and the species can spread to a wider range of habitats. In general, this will lead to errors of omission in the predictions.

In performing this project we have identified that this is only one of the reasons that predictive performance is poor. Another reason is that the statistical relationships between distribution and environmental variables may be different in the native and invaded range. This effect is driven by the gap between the real processes driving distribution and the observed summary environmental data available for modelling. These relationships vary spatially leading to failures in predictions to new environments. This effect can be potentially exacerbated by over fitting. We can make models fit more and more precisely in the native range but this does not necessarily increase the quality of the predictions in the invaded range. In the following chapter we discuss further our decision to restrict the model fitting to the native range only.

In this chapter we explore this issue by using simulation. This will provide a clearer understanding of the challenges in prediction, and allows consideration of new approaches.

Methods

To investigate the effect of incomplete process knowledge we seek to isolate this phenomenon from other effects such as biotic interactions and biased sampling. If this is not done it will be more difficult to interpret the results of the analysis. With this aim in mind it is clear that the use of observed data is not optimal for this investigation – it explicitly confounds these competing effects.

Instead we consider simulating data. To do this we begin by considering the logical basis of the analysis. The methods all assume that there is a “niche” within which the species occurs. Once this niche is defined it determines the presence and absence of the species irrespective of location. To mimic this we choose the simplest expression of it. We choose a temperature variable from the BIOCLIM set (1-11) and a moisture variable (12-19). We set limits for each of these variables. Within these limits (a box in the environmental space of these two variables) the species is always present. Outside of the box the species is always absent. Thus these synthetic species exhibit pure niches, with no biotic interactions, and they have no stochastic element, i.e. the probability of presence is zero or one. The limits were set based on a random selection across the global range of the environments. These “niches” were then realised in South America and Australia. Thus some simulated species might be rare in one of these continents and common elsewhere.

The simulations in this analysis consider the potential native range to be locations within the niche in South America and the potential invaded range to be anywhere in Australia. Given we have gridded data for the

BIOCLIM variables across this range we can produce the geographic distribution exactly based on the environmental box defined in the previous paragraph. The important point to note is that we have “truth” in both localities so we can unambiguously assess performance.

The issue that we wish to explore is the effect of incomplete process knowledge on our ability to predict the invaded distribution. In practical terms it means that our predictor variables are often not close to process. For example a species does not experience mean temperature, but rather a changing profile of temperature minute by minute, the impact of which interacts with other environmental variables and the species physiology and behaviour. We mimic this by predicting the species distribution not by the variables that define its niche, but by the remaining BIOCLIM variables. These variables will be correlated with the “causal” variable mimicking the usual problem in extrapolating these models. Note that the true variables are NOT available to the model selection process. This is to explicitly consider the impact of less proximal variables.

Based on these models there are three comparisons that are of interest:

1. Build model in native range and project into both native and invaded range.
2. Build model in invaded range and projected into invaded range
3. Build model in native range, build model in invaded range, based on the same data and compare predictions.

For the purpose of this study we use generalised additive models with smoothness degrees of freedom set to 4. These models will adequately fit the “true” distribution and will not unduly impact the results of the analysis. The response is presence-absence. Predictors are selected as before – i.e. the pair of temperature and rainfall predictors leading to smallest residual deviance are chosen. Predictions are assessed using area under the receiver operating characteristic (ROC) curve. This ranges from 0 to 1, with random predictions giving a ROC of 0.5, and ROC values reporting the proportion of times that a prediction at a randomly drawn suitable (presence) site will be greater than the prediction at a randomly drawn unsuitable (absence) site.

Results

Individual models

We simulated over one thousand different distributions to explore extrapolation performance. To fix ideas we initially present two models chosen which show different extremes of performance. One performs reasonably the other less so. In the first model the predictions are particularly poor. The South American data is in significant spatial disequilibrium with the environmental covariates (i.e. the spatial predictions into the native range have substantial mismatches with the observed truth), even though each was found to be highly significant and the ROC >.75. In the second example the model performs significantly better as the observed covariates are good surrogates for the underlying proximal covariates.

Example 1:

True niche – BIOCLIM 2 and BIOCLIM 19

Empirical niche – BIOCLIM 4 and BIOCLIM 14 selected as best fit.

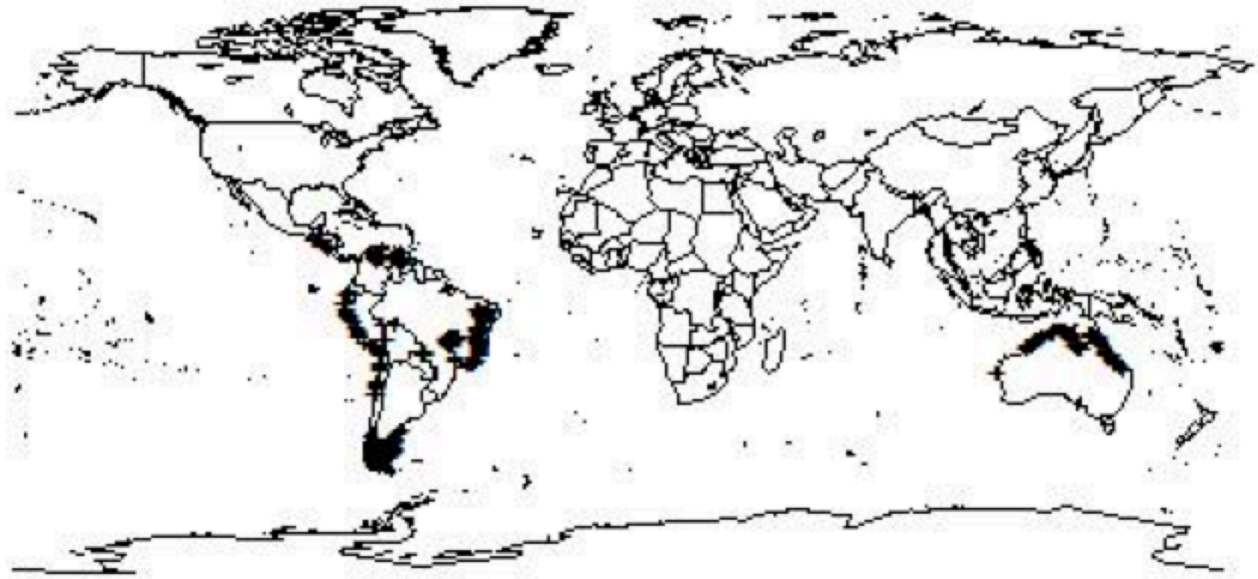


Figure 33. Simulated distribution, BIOCLIM variables 2 and 19.

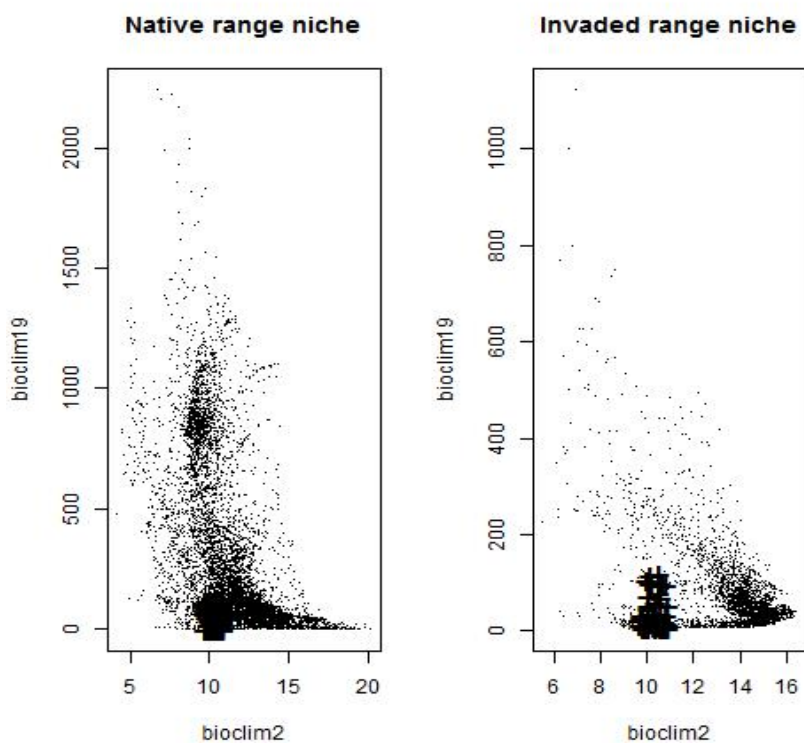


Figure 34 Actual niche. The block of + symbols show the presences, and small dots are absences. Note the axes are not scaled identically in the two panels.

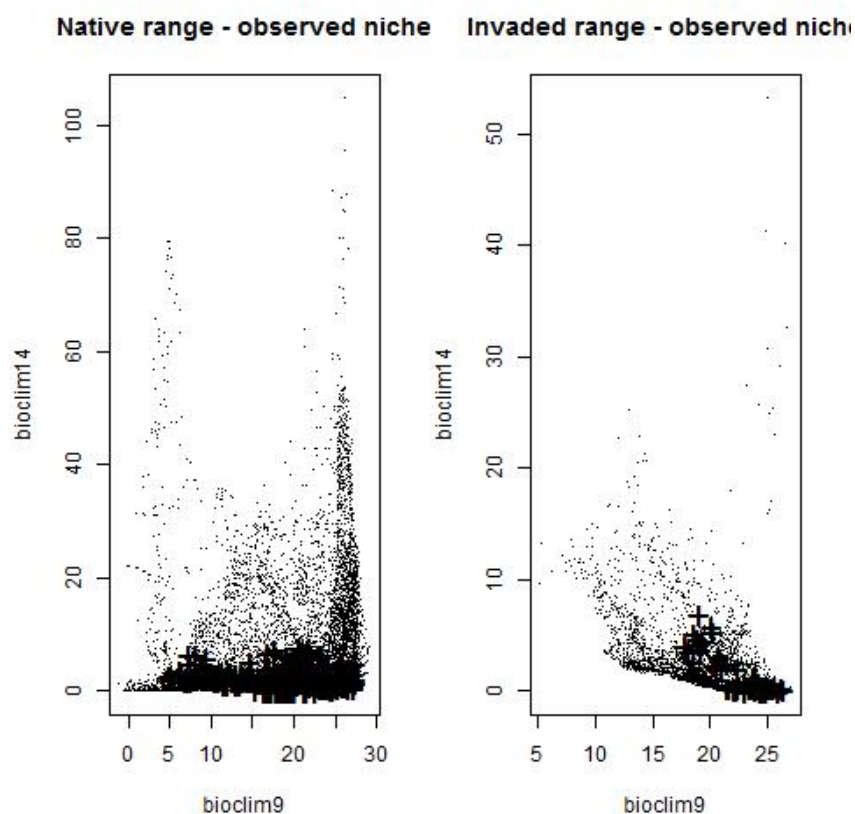


Figure 35 Empirical niche – observations shown along the selected variables, 9 and 14. Legend as above.

p predicted from home range model vs from introduced range mod

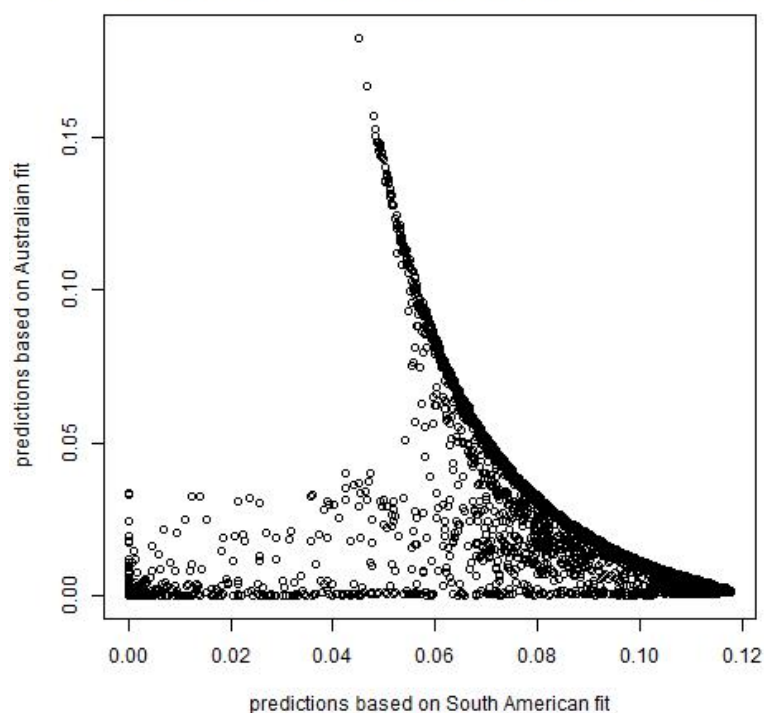


Figure 36. Comparison of predictions from model fitted to South American data to model fitted to Australian data, to all environments on the Australian continent.

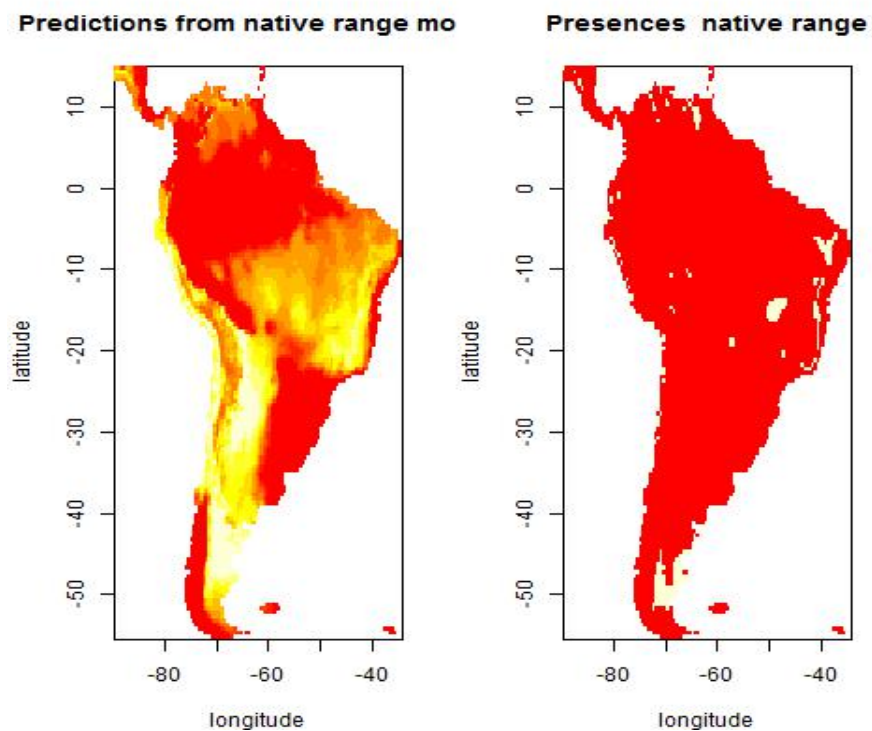


Figure 37 Predictions and actual data for South American data .
 Note: legend for this and following: values are high (pale yellow) to low (red)

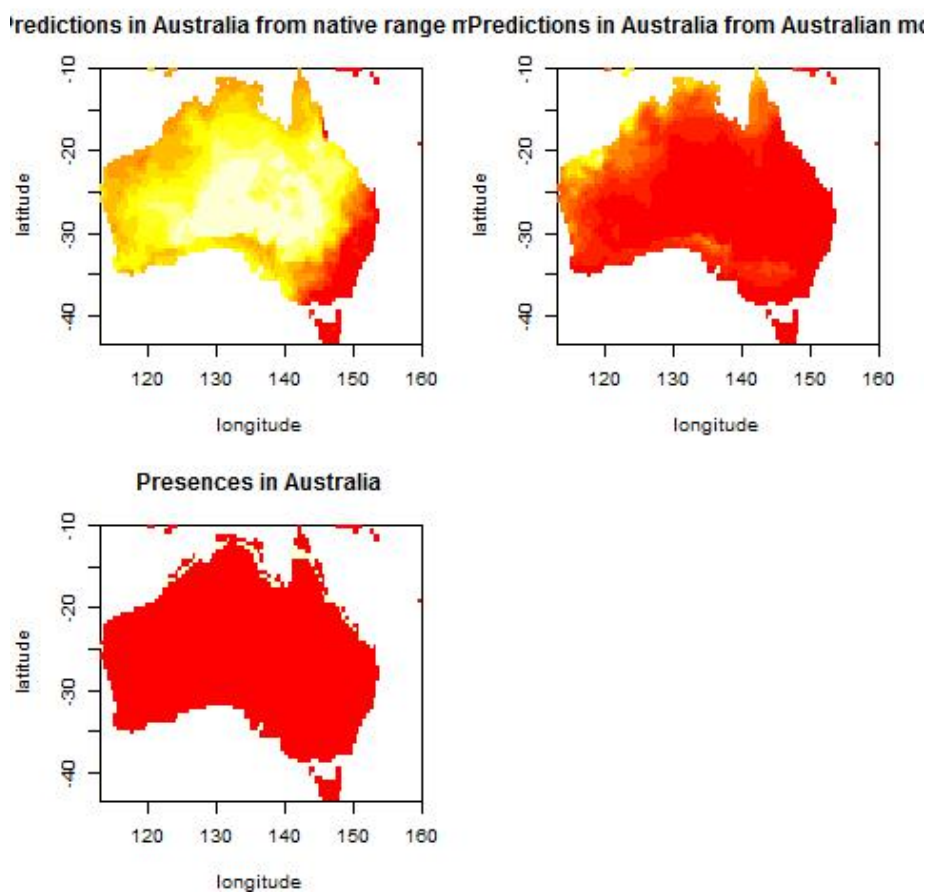


Figure 38. Predictions from native range model, Australian model and actual data.

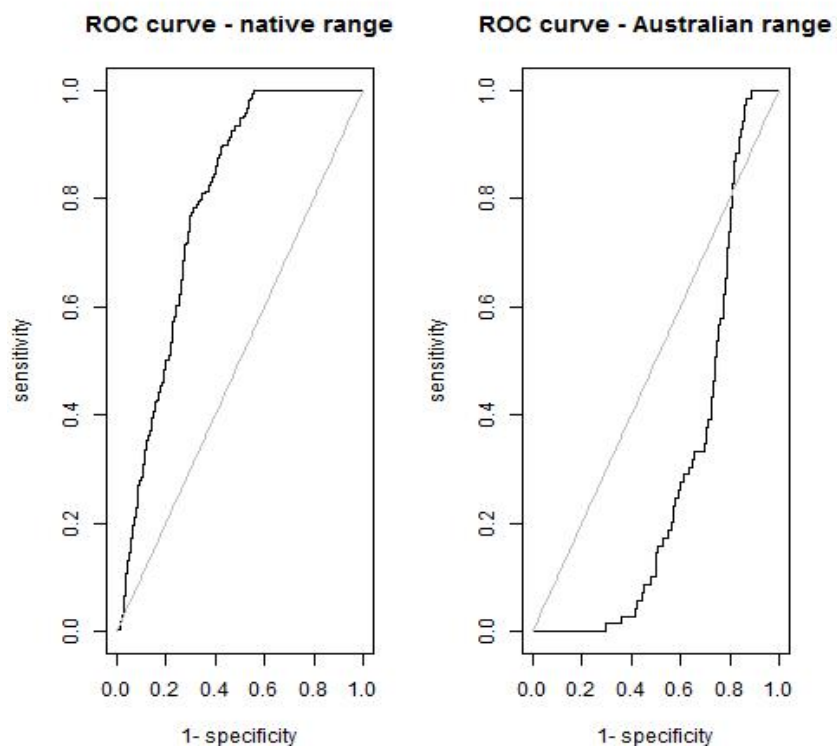


Figure 39. ROC curves from native range model and for projections to Australia.

Example 2:

True niche – BIOCLIM 11 and BIOCLIM 12

Empirical niche – BIOCLIM 5 and BIOCLIM 16

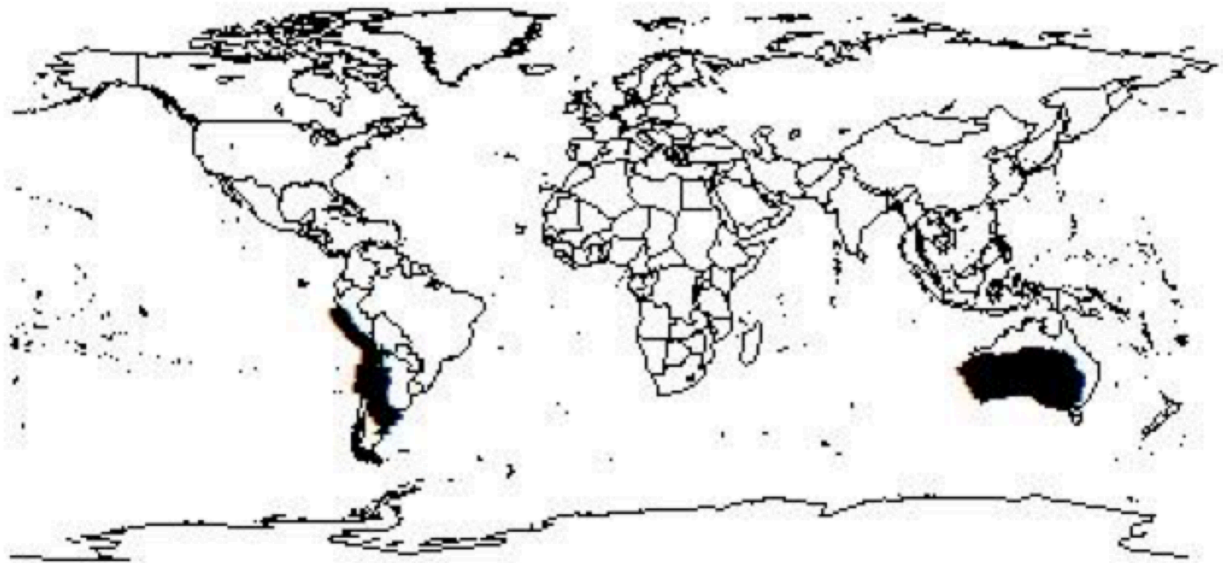


Figure 40. Simulated distribution, BIOCLIM variables 2 and 19.

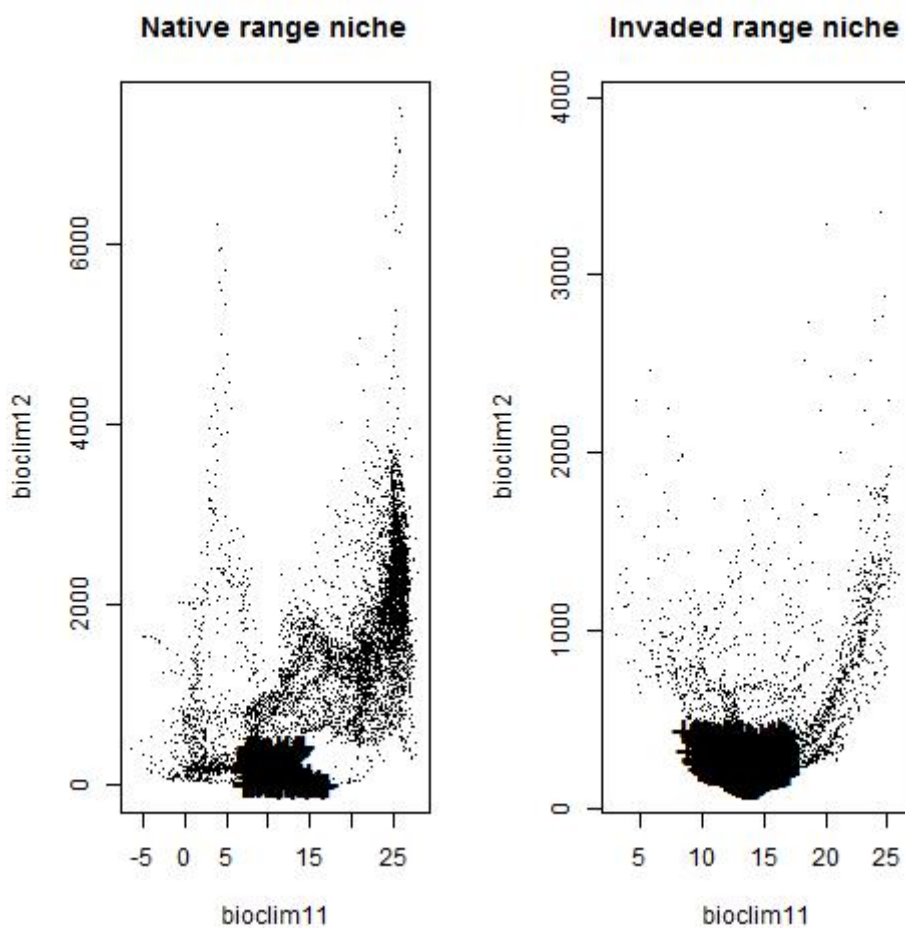


Figure 41. Actual niche.

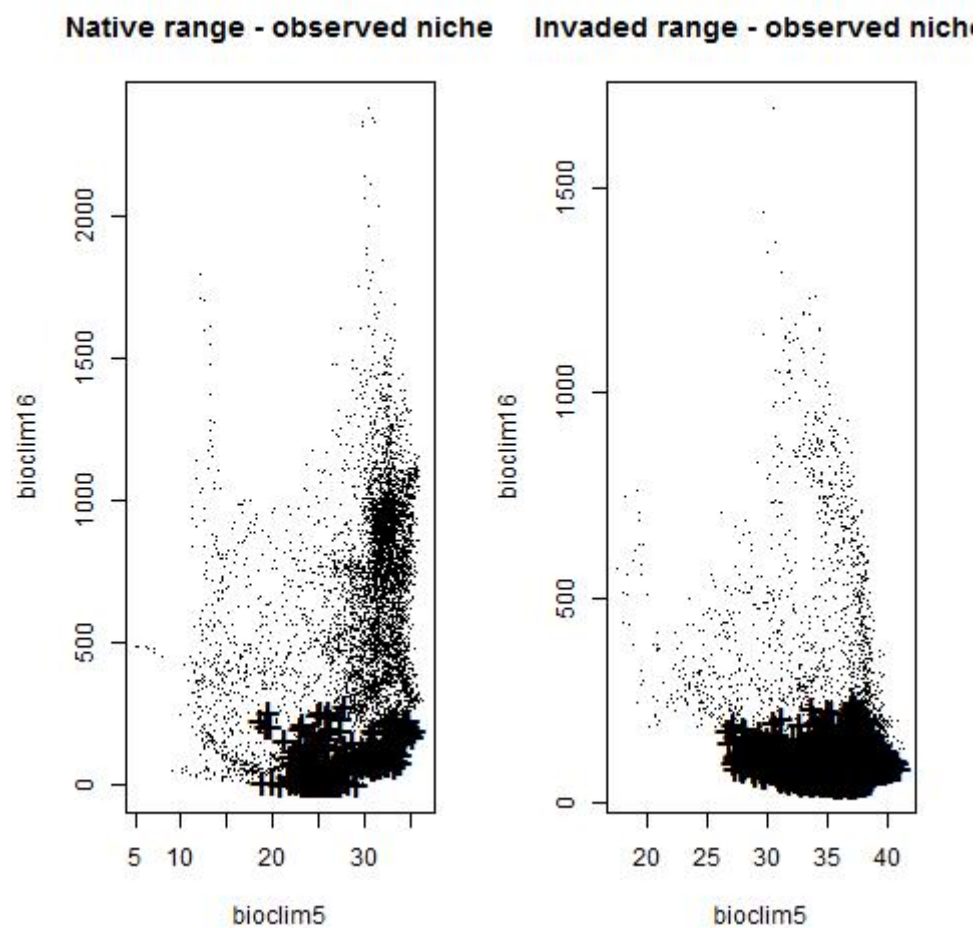


Figure 42. Empirical niche.

p predicted from home range model vs from introduced range mod

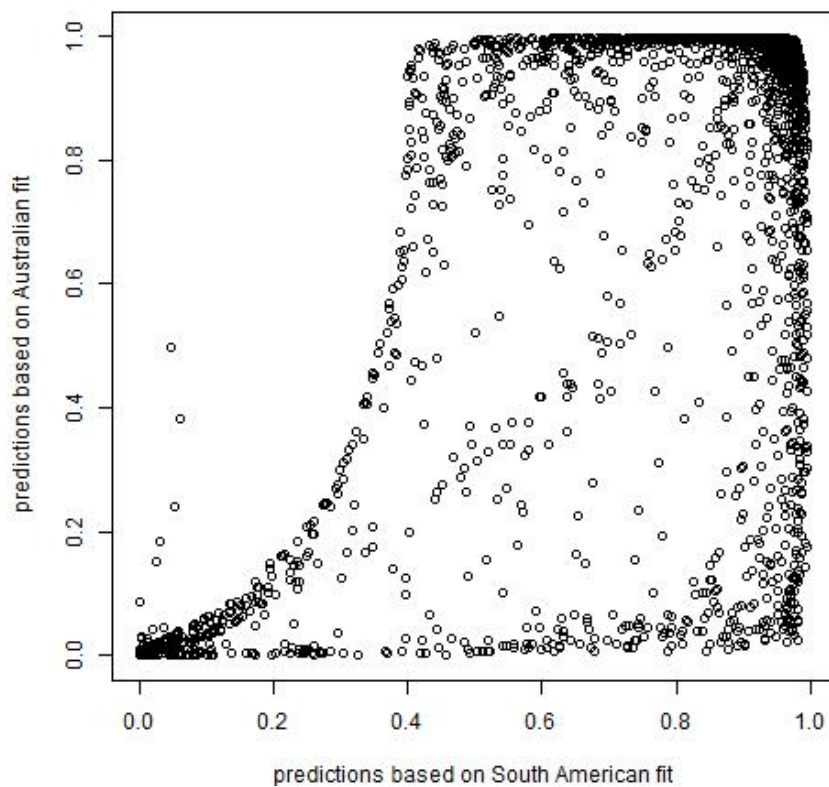


Figure 43. Comparison of predictions from model fitted to South American data to model fitted to Australian data. Predictions based on Australian environmental data.

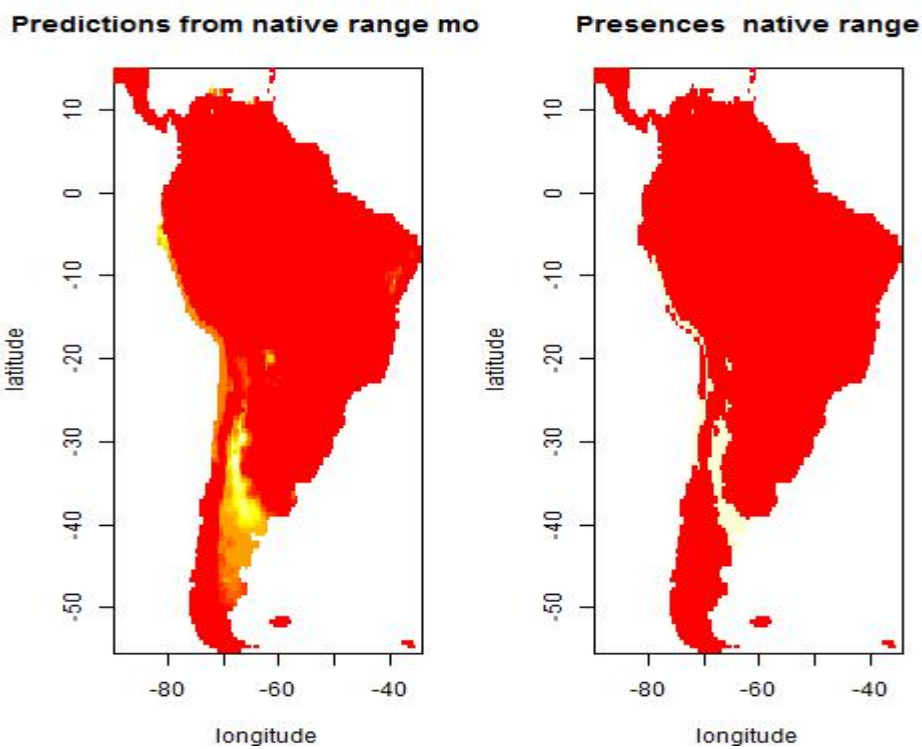


Figure 44. Predictions and actual data for South American data.

Predictions in Australia from native range model Predictions in Australia from Australian model

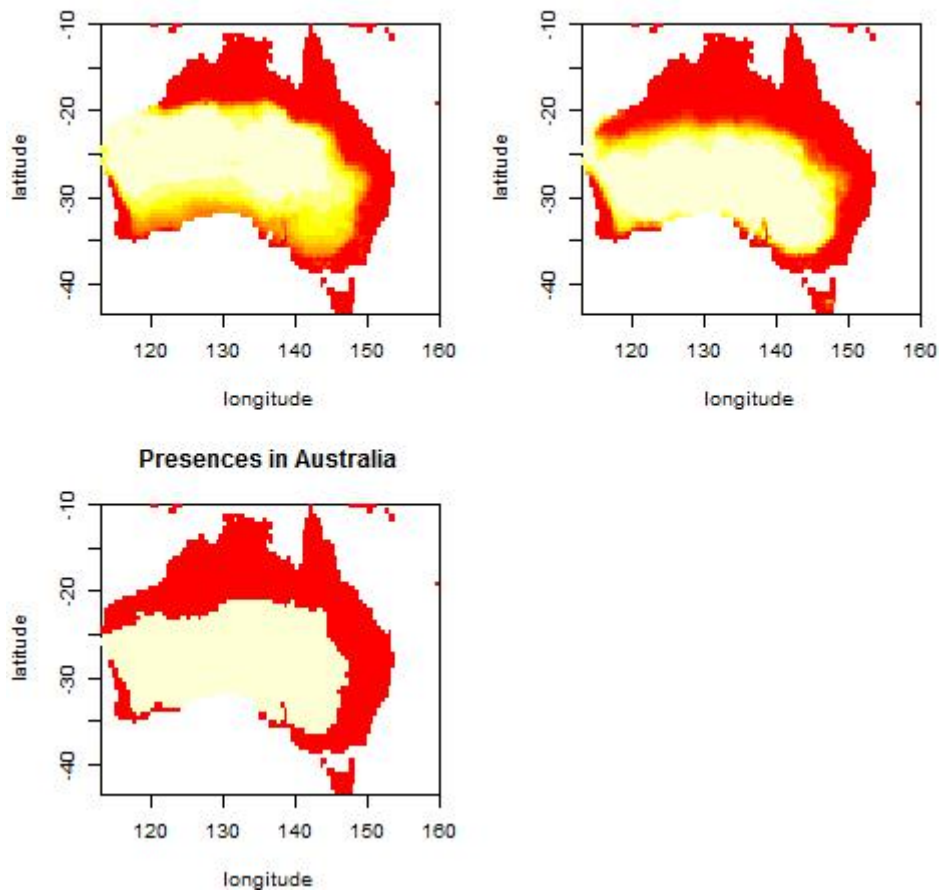


Figure 45. Predictions from native range model, Australian model and actual data.

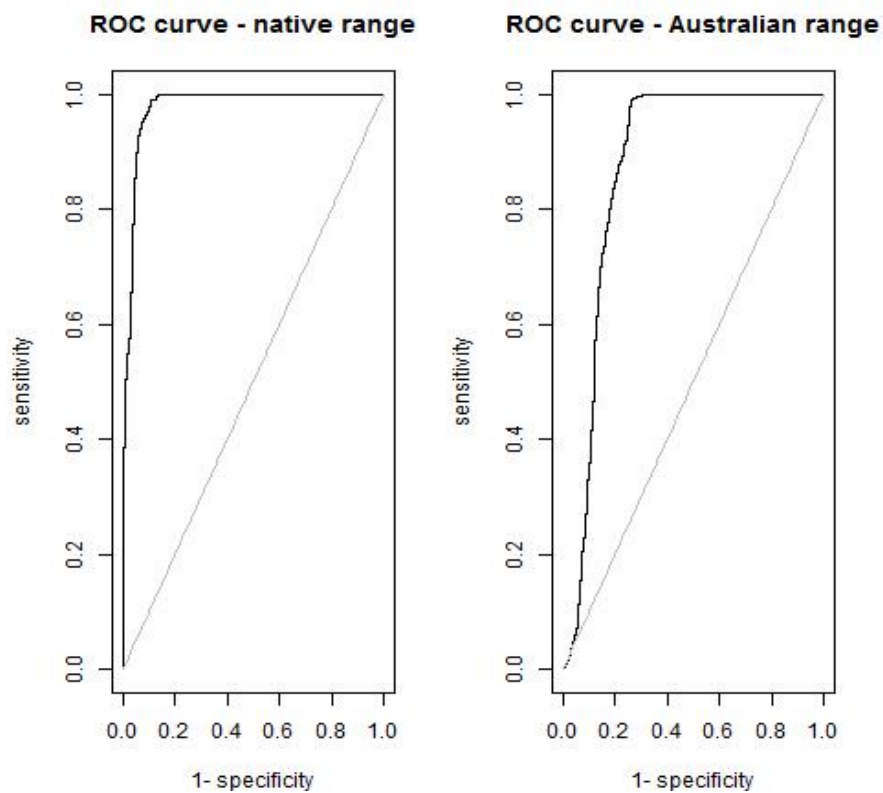


Figure 46. ROC curves from native range model and for projections to Australia.

These two examples highlight the challenges of predicting species distributions to new environments based on limited environmental data. While a niche may be well defined and universally applicable, for example Figure 40, this ceases to be true for the relationship based on surrogate variables. For an example compare Figures 40 and 41. This is the central issue and in practical applications the “truth” is unknown. The impact of this depends in the extent of degradation in relationships caused by the use of surrogate variables. In the first example the effect is extreme to the extent that the model would be misleading for decision making. In the second case the impact is less pronounced.

In either case the impact on the probability predictions is significant, to the extent that it would seem particularly unwise to apply these in decision making. Even without the results from this simulation it is easy to see that the two regions are different statistical populations and population statistics in one would be only loosely related to population statistics in the other. In other words, the prevalence of the species (the proportion of the land area occupied) is not the same on each continent. A reasonable question to ask is whether the presented results are not representative. To explore this we have rerun the simulation for random pairs of temperature and moisture variables. Using the same code for each random pair we have recorded the performance, in terms of AUC in the home range against the performance in the invaded range.

We present the results of this simulation in Figure 46. Note in this figure that all models have an AUC > .75 so would at least be anecdotally be considered reasonable. Note also the significant degradation that occurs in many models performance when they are projected, some of it potentially extreme. In particular note that predictive performance in the invaded range is in practical terms uncorrelated to predictive performance in the native range.

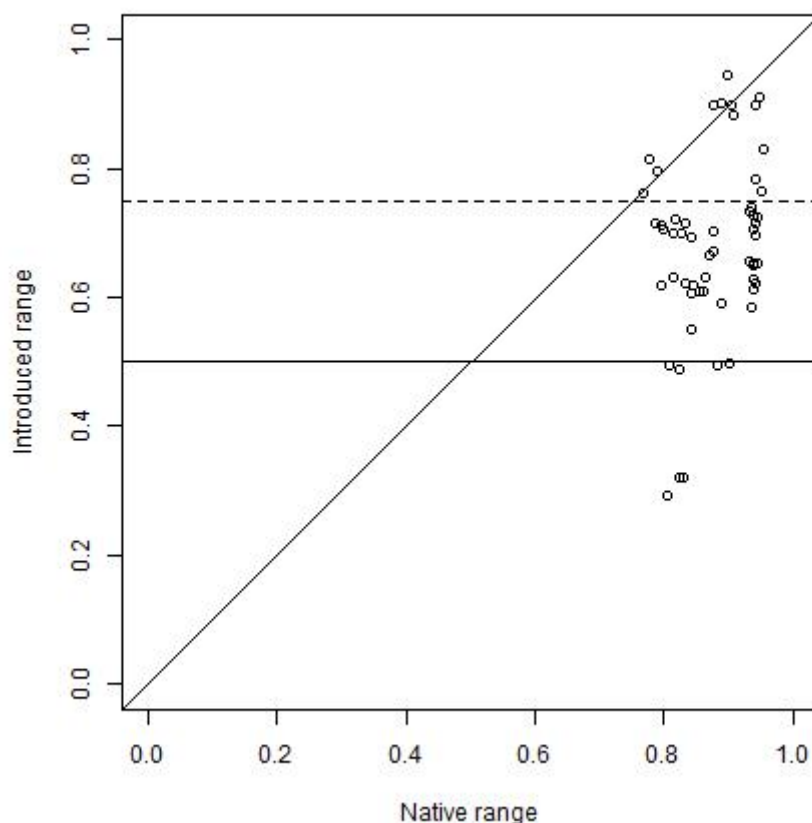


Figure 47. 100 randomly sampled models, ROC curve based on fit and projection to Australia.

If we cannot identify models that will project well based on empirical performance perhaps there will be particular variables that are associated with good projection properties. To explore this we sampled 100 random niches, randomising both the variables included (but constraining them to one BIOCLIM temperature and precipitation variable) and the position of the niche along the gradient. For each generated niche we fitted all 88 possible BIOCLIM rainfall x temperature combinations to the native range. For each model we recorded the AUC of the projection onto the Australian data. These were averaged over the 100 runs to produce average AUC's for each variable combination. This is shown in Figure 47. Note that there are no obvious patterns in this Figure.

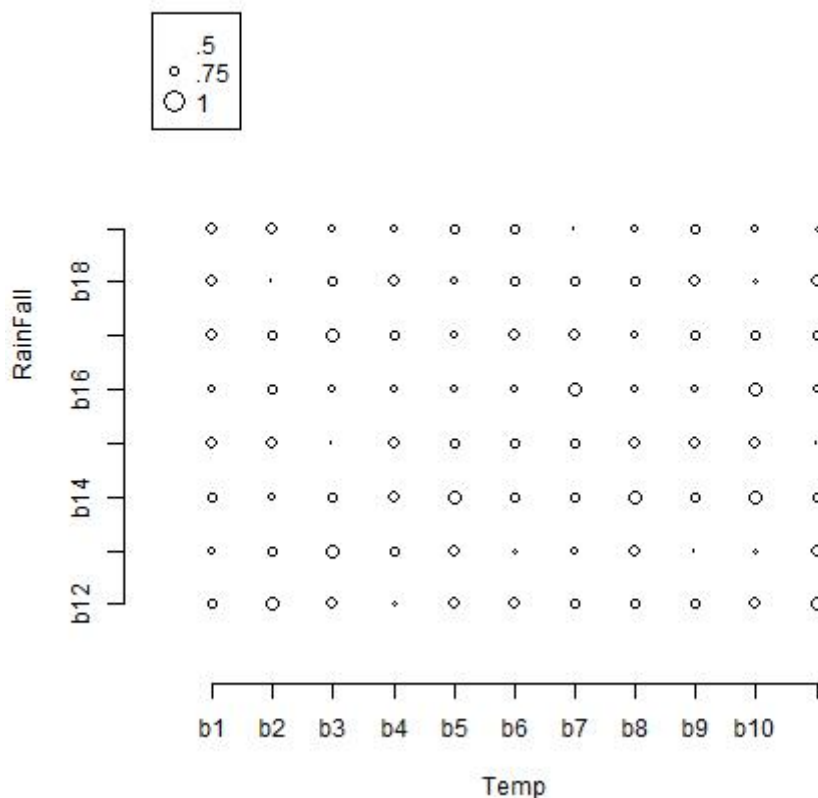


Figure 48 100 randomly sampled models, ROC curve based on comparison of projection to Australia against truth observations in Australia.

Conclusions

This chapter has identified the significant challenges for prediction when dealing with incomplete information. Many models will appear to fit due to the spatial nature of environmental variables but the projections of these will often be unreliable. There are two approaches to dealing with this. First, more detailed laboratory studies can be performed to better understand physiological tolerances. This is potentially complicated by genetic variation across a species as well as the more pernicious issue that the mechanisms controlling distribution may be multi-factorial. In this case it may not be possible to design effective experiments to identify these issues. In addition, the time frames of many Biosecurity questions are short – weeks and months, rather than years. It may not be possible to collect information in this timeframe. The second approach is to accept this uncertainty and to factor it into the inference. We will discuss this in the next chapter.

7 Synthesis

7.1 Discussion

We have reviewed over 40 papers in the ecology literature related to variable selection in prediction and consulted the statistical literature to consider the impacts of dimensionality on distance based predictions. From this review we have concluded that there is no accepted position about what constitutes proximal variables that will consistently predict accurately to new locations. There is some good conceptual thinking for certain classes of organisms but no general agreement on how to translate this into a choice between the set of available variables.

The review of this literature identified a number of areas of agreement about predicting the range of invasive species. These are:

- There are fundamental issues with predicting to new environments because of biotic interactions. These are generally viewed as irreducible.
- There is conceptual agreement that some variables are more predictive than others (the proximal/distal debate) but there is no agreement about a constructive approach to identify these variables. There is broad agreement that some variables can be identified as NOT being proximal but there is no general agreement about determining the ranking of variables that are proximal.

These results are somewhat sobering. While the concept of proximal variables is often invoked (along with the allied “expert ecological knowledge”), the literature remains conceptual without a developed empirical basis and synthesis. While reasonable predictions can be made in some well-studied examples the absence of a general theory means that empirical justification can be on tenuous logical ground (*cum hoc ergo propter hoc* - “with this, therefore because of this”). The more “causal” variables we consider the more likely we will find association (in the statistical sense of reasonable model fit), but as we have demonstrated, these associations may not reflect actual process and may therefore not project well.

The inconsistent results obtained from the analysis of statistical approaches to identifying proximal variables, using the five case study species, are also a cause for reflection. Proximal variables should be predictive. The results show that we have difficulty in predicting in ways are consistent with current knowledge / opinion on these species. Reasons may include: there could be strong biotic interactions; dispersal limitations are strongly impacting distributions; the variables that are available to us are only weakly related to process; or the data we have available for fitting the models is too poorly representative of the species distribution to enable strong model fit. After contemplation this is perhaps not surprising. The variables that are widely available to modellers are often coarse and/or abstract averages across significant temporal and spatial scales and the processes that drive distributions relate to individual level interactions that will occur at particular places in space and time. Thus the conceptualisation of a niche is logically significantly removed from the mechanics of the modelling process and these difficulties in projection should be anticipated.

This project sought to identify best practice approaches to developing distribution predictions. The logical conclusion from the review and analysis of case studies is that a clear consensus on good practice does not exist at this stage and that any process using existing methods will involve a significant expert component. Without an objective method to determine variables there is no other way. The expert(s) are needed to justify particular choices about possible causal relationships. This is still challenging – experts learn by observation so their understanding of causation is built on a foundation of correlation. We learn by observing correlations and building conceptual models to rationalise them. There will therefore inevitably be errors within these conceptual models but provided the experts are seen as reliable by relevant

stakeholders and processes have been put in place to minimise bias they still represent a constructive way forward.

The reliance on experts could cause difficulties in contentious policy areas. If the only differentiator between logical possibilities is the “expert”, competing interest can engage their own experts that espouse views consistent with the party engaging the expert. While this is entirely reasonable in a contentious debate it opens up questions about what is expertise in this context. It also introduces a moral hazard – the lack of ability to resolve a question introduces the possibility of parties choosing positions based on favoured outcomes rather than reason. Given humans capacity to confuse correlation with causation significant challenges may arise. An additional problem occurs when decisions are time critical. In this case extended debate between experts on topics that cannot be empirically resolved can lead to unacceptable delays.

The implication of this is that groups with relevant expertise need to be convened to undertake these analyses. When the results are time critical, such as in emergency response, the approach used to manage/resolve variations in expert opinions need to be codified so that it can be applied efficiently to support timely decision making. This will typically mean that it needs to be agreed at the policy level and then applied as needed. When analyses are not as time critical greater flexibility in process can be entertained.

In convening these experts it is important to provide sufficient opportunity for them to express their views and have them constructively challenged as appropriate. To this end the project has compiled a range of data sets, and tools that can be used to generate potential predictive variables as needed. These should form the basis of a data library that can support these activities in the future and be added to as new datasets are developed or identified. In particular these will provide the basis of a more systematic approach to these problems.

While there are major challenges in assessing the predictive performance of models in new locations it is still useful to model and analyse the distribution of the species in their native range. Hypotheses about causal variables can still be assessed. Distributions of species in environmental space can be considered and strong disjunctions explored. Different possibilities can be examined. But the experts must not rely purely on predictive performance to guide a single choice. As we have demonstrated, this is a flawed approach as fit in the home range is not a good predictor of extrapolation performance.

While we have demonstrated that a single approach cannot be recommended to perform in all circumstances there are other general conclusions that can be made. In particular our analysis has identified a number of approaches that should not be considered by experts. The use of statistical models to produce probability-based predictions in invaded ranges should be used extremely cautiously. As we have argued in this report, the probabilities are population specific and the dangers of extrapolating from one population to another are well known. In particular the probabilities should never be interpreted as reflecting the expected proportion of sites that will be invaded as this will typically have no logical foundation. One could argue on the basis of “all things being equal” that the probabilities may be useful in a relative sense but this relies as much on faith as on logic. As an alternative we recommend that practitioners use enveloping methods such as alpha hulls to define the regions of environmental space that are potentially inhabitable by the organism.

We also strongly recommend against over-fitting models in terms of including a large number of variables, particularly if these are chosen based on availability rather than physiology. Specialisation of models to the native range will typically not lead to better projections as the key process driver are unknown and differentially related to the presences across the two locations. Smaller models are also more conservative in the sense that they will predict over larger areas in the invaded range. It may be possible that models with more than the two variables considered here would improve performance in some circumstances. The difficulty with this is that there is no empirical way of identifying this. Again experts must be used to determine this.

Extrapolation into regions with novel climates, in the sense that the climate does not exist in the native range should also be considered carefully. Tools such as MESS (Elith et al 2010) should be used to identify

these regions but its general utility in this context needs to be tested. How this enters into decision making will depend on the context and the nature of the novelty.

The choice of background points is not clear cut. Based on the case studies the use of continental background was favoured. There is still not a clear logic to any particular choice. Background points that too closely follow the observed distribution may miss variables associated with broader scale distribution differentiation. These may be important for continental predictions into Australia/NZ. But background points selected on a global basis are clearly not relevant as they significantly confound dispersal barriers with abiotic and biotic factors.

7.2 Protocol

The protocol we propose attempts to address the issues we have identified. The key points for the Australian and NZ governments in decision making are:

- The lack of systematic approaches in the literature means that an ad-hoc approach to the species prediction problem risks techniques that will not survive significant scrutiny in contested decision making.
- A smaller set of variables that are more process relevant will provide a more defensible prediction and a greater test of experts understanding than approaches that use large numbers of variables and attempt to automate the analysis.

The protocol that we propose is systematic and will develop an appropriate knowledge base and code of practice over time.

1. If detailed, well-supported physiological information exists, it should be used to make projections. In particular the physiological information should be used as the basis for correlative modelling. If detailed physiological information does not exist, experts (including organism experts AND distribution modelling experts) should be convened to identify possible proximal sets of variables and assess these by considering the correlative evidence from the native range. The expert process should be facilitated to ensure that uncertainty about possible “proximal” predictors is identified in the analysis and carried forward. Thus they need to consider multiple sets of variables. Experts should be beware of over specialising models and as a starting point include one temperature and one moisture related variable.
2. The observed distribution data in the native range and each set of variables should be used to construct alpha hulls if the number of variables is 3 or less. Currently, for dimension greater than three there is no readily available code (in R) to assess whether points are within convex hulls or alpha hulls.
3. An analysis using techniques such as MESS should be performed to identify any locations (in terms of environment) in the invaded locations not represented in the training data.
4. Results should be presented for each predictor set identified by experts.

An important point in the protocol is that uncertainty around predictors is carried forward in the analysis to the decision phase. At this point choices about the use of best/worst or most likely case can be done on a policy basis depending on the context of the decision. For decisions that need to be made rapidly, such as assessing costs/benefits of eradication, the process should be codified to ensure that models are implemented consistently and decisions are made efficiently.

7.3 Future developments

This study has identified a number of possible avenues of future research and synthesis.

A key result of this study is the identification of the lack of relevant theory regarding the choice of proximal sets of predictors. There is a need to further develop the conceptualisation of the processes that determine the distribution of species. There is also the opportunity to develop an empirical understanding of what factors are associated with the observed distributions of different groups of organisms. New thinking and a greater body of empirical evidence is needed to make progress on this topic. This study has identified that there may be fundamental limits on our ability to predict species performance in novel climates.

Here we focussed on fitting models to records from the native range only, but some authors argue for inclusion of records from invaded ranges where the species is long established (Chapter 2). We recognise that there are valid arguments for this viewpoint, and testing whether inclusion of more records improved performance in our approaches would be worthwhile.

This study has highlighted the inherent uncertainty that is involved in predicting species distributions in new environments. The response to this has been to explicitly incorporate this into the protocol to ensure that this uncertainty is carried through for consideration by decision makers. We anticipate that further work needs to be done to determine ways of incorporating this efficiently in the decision making process. Uncertainty on its own can impede decision making so it is important that this issue is addressed. We note that different decisions could require different approaches to dealing with uncertainty. Techniques such as model averaging could be useful to provide intermediate products that can be used directly in decision making.

The study has posed a number of new questions about predicting invasive species distributions. In Chapter 6 we have briefly explored whether some variables are on average more predictive than others in projecting from one location to another. This could potentially be investigated more broadly based on more extensive information about native and invaded ranges rather than the simulated data used here.

The central role of experts in this process suggests that better tools to diagnose/estimate potential limiting factors would be useful to support the experts' discussions about proximal variables. While noting that this project has highlighted the challenge of assessing causations, it is still important to provide experts with as many tools as possible to assess hypotheses and validate/invalidate observations.

The final point to consider is that in many situations, knowledge of a species ecology and physiology will be limited. Potential invaders can be difficult to predict and by definition come from locations outside of Australia/NZ so practical experience about the species is often extremely limited. Thus we anticipate the need to develop a default distribution of sets of potential variables that can be used in the analysis. The choice of these variables needs to be considered further. At this stage the BIOCLIM set are used in a majority of studies as a starting point. But the context here is different. Current studies select a subset of these variables on the basis of correlation with distribution. We are interested in representing uncertainty across possible proximal sets. Thus any final choice will require additional analysis and policy input. There is the opportunity to investigate development of variables following the work of Sutherst and Maywald (1985). In particular exploring approaches to statistical parameter estimation for simple models could be considered.

8 References

- Aalto, J., le Roux, P., and Luoto, M. (2014). The meso-scale drivers of temperature extremes in high-latitude Fennoscandia. *Climate Dynamics* **42**, 237-252. doi: 10.1007/s00382-012-1590-y.
- Allen, J. C., Foltz, J. L., Dixon, W. N., Liebhold, A. M., Colbert, J. J., Regniere, J., Gray, D. R., Wilder, J. W., and Christie, I. (1993). Will the gypsy moth become a pest in Florida? *The Florida Entomologist* **76**, 102-113. doi: 10.2307/3496018.
- Ashcroft, M., Gollan, J., and Batley, M. (2012a). Combining citizen science, bioclimatic envelope models and observed habitat preferences to determine the distribution of an inconspicuous, recently detected introduced bee (*Halictus smaragdulus* Vachal Hymenoptera: Halictidae) in Australia. *Biological Invasions* **14**, 515-527. doi: 10.1007/s10530-011-0092-x.
- Ashcroft, M. B., French, K. O., and Chisholm, L. A. (2011). An evaluation of environmental factors affecting species distributions. *Ecological Modelling* **222**, 524-531. doi: <http://dx.doi.org/10.1016/j.ecolmodel.2010.10.003>.
- Ashcroft, M. B., Gollan, J. R., and Batley, M. (2012b). Combining citizen science, bioclimatic envelope models and observed habitat preferences to determine the distribution of an inconspicuous, recently detected introduced bee (*Halictus smaragdulus* Vachal Hymenoptera: Halictidae) in Australia. *Biological Invasions* **14**, 515–527. doi: 10.1007/s10530-011-0092-x.
- Austin, M. P. (1980). Searching for a model for use in vegetation analysis. *Vegetatio* **42**, 11-21. doi: 10.1007/bf00048865.
- Austin, M. P. (2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling* **157**, 101-118. doi: [http://dx.doi.org/10.1016/S0304-3800\(02\)00205-3](http://dx.doi.org/10.1016/S0304-3800(02)00205-3).
- Austin, M. P. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling* **200**, 1-19.
- Austin, M. P. and Smith, T. M. (1989). A new model for the continuum concept. *Vegetatio* **83**, 35-47. doi: 10.1007/bf00031679.
- Austin, M. P. and Van Niel, K. P. (2011a). Impact of landscape predictors on climate change modelling of species distributions: a case study with *Eucalyptus fastigata* in southern New South Wales, Australia. *Journal of Biogeography* **38**, 9-19. doi: 10.1111/j.1365-2699.2010.02415.x.
- Austin, M. P. and Van Niel, K. P. (2011b). Improving species distribution models for climate change studies: variable selection and scale. *Journal of Biogeography* **38**, 1-8. doi: 10.1111/j.1365-2699.2010.02416.x.

- Barbet-Massin, M. and Jetz, W. (2014). A 40-year, continent-wide, multispecies assessment of relevant climate predictors for species distribution modelling. *Diversity and Distributions* **20**, 1285-1295. doi: 10.1111/ddi.12229.
- Bertelsmeir, C. and Courchamp, F. (2014). Future ant invasions in France. *Environmental Conservation* **41**, 217-228. doi: 10.1017/S0376892913000556.
- Besnard, A. G., La Jeunesse, I., Pays, O., and Secondi, J. (2013). Topographic wetness index predicts the occurrence of bird species in floodplains. *Diversity and Distributions* **19**, 955-963. doi: 10.1111/ddi.12047.
- Biosecurity Australia (2009). Draft report: pest risk analysis report for Guava Rust. Biosecurity Australia. (Canberra.)
- Bradley, B. A., Olsson, A. D., Wang, O., Dickson, B. G., Pelech, L., Sesnie, S. E., and Zachmann, L. J. (2012). Species detection vs. habitat suitability: Are we biasing habitat suitability models with remotely sensed data? *Ecological Modelling* **244**, 57-64. doi: 10.1016/j.ecolmodel.2012.06.019.
- Braunisch, V., Coppes, J., Arlettaz, R., Suchant, R., Schmid, H., and Bollmann, K. (2013). Selecting from correlated climate variables: a major source of uncertainty for predicting species distributions under climate change. *Ecography*, no-no. doi: 10.1111/j.1600-0587.2013.00138.x.
- Broennimann, O. and Guisan, A. (2008). Predicting current and future biological invasions: both native and invaded ranges matter. *Biology Letters* **4**, 585-589. doi: 10.1098/rsbl.2008.0254.
- Bucklin, D. N., Basille, M., Benscoter, A. M., Brandt, L. A., Mazzotti, F. J., Romañach, S. S., Speroterra, C., and Watling, J. I. (2015). Comparing species distribution models constructed with different subsets of environmental predictors. *Diversity and Distributions* **21**, 23-35. doi: 10.1111/ddi.12247.
- Cantrell, B., Chadwick, B., and Cahill, A. (2002) 'Fruit fly fighters: eradication of the papaya fruit fly.' (CSIRO PUBLISHING: Melbourne.)
- Carnegie, A. J. and Cooper, K. (2011). Emergency response to the incursion of an exotic myrtaceous rust in Australia. *Australasian Plant Pathology* **40**, 346-359.
- Carroll, J. and Marks, M. (2003). Asian Gypsy Moth. USDA APHIS Factsheet.
- Chatfield, B. S., Van Niel, K. P., Kendrick, G. A., and Harvey, E. S. (2010). Combining environmental gradients to explain and predict the structure of demersal fish distributions. *Journal of Biogeography* **37**, 593-605. doi: 10.1111/j.1365-2699.2009.02246.x.
- Commonwealth of Australia (2015). National pests & disease outbreaks - Red Imported Fire Ants. (Commonwealth of Australia.)
- Coudun, C., Gegout, J. C., Piedallu, C., and Rameau, J. C. (2006). Soil nutritional factors improve models of plant species distribution: an illustration with *Acer campestre* (L.) in France. *Journal of Biogeography* **33**, 1750-1763. doi: 10.1111/j.1365-2699.2005.01443.x.
- CSIRO (2015). Ants Down Under - *Solenopsis invicta* Buren, 1972. (CSIRO.)

DAFF (2015). Fire ants biology and ecology. (Queensland Government Department of Agriculture and Fisheries.)

De Meyer, M., Robertson, M. P., Mansell, M. W., Ekesi, S., Tsuruta, K., Mwaiko, W., Vayssieres, J. F., and Peterson, A. T. (2010). Ecological niche and potential geographic distribution of the invasive fruit fly *Bactrocera invadens* (Diptera, Tephritidae). *Bulletin of entomological research* **100**, 35-48. doi: 10.1017/s0007485309006713.

Diniz-Filho, J. A. F., De Marco Jr, P., and Hawkins, B. A. (2010). Defying the curse of ignorance: perspectives in insect macroecology and conservation biogeography. *Insect Conservation and Diversity* **3**, 172-179. doi: 10.1111/j.1752-4598.2010.00091.x.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carre, G., Garcia Marquez, J. R., Gruber, B., Lafourcade, B., Leita, P. J., Muenkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schroeder, B., Skidmore, A. K., Zurell, D., and Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **36**, 27-46. doi: 10.1111/j.1600-0587.2012.07348.x.

Dubuis, A., Giovanettina, S., Pellissier, L., Pottier, J., Vittoz, P., and Guisan, A. (2013). Improving the prediction of plant species distribution and community composition by adding edaphic to topo-climatic variables. *Journal of Vegetation Science* **24**, 593-606. doi: 10.1111/jvs.12002.

Dungan, J. L., Perry, J. N., Dale, M. R. T., Legendre, P., Citron-Pousty, S., Fortin, M. J., Jakomulska, A., Miriti, M., and Rosenberg, M. S. (2002). A balanced view of scale in spatial statistical analysis. *Ecography* **25**, 626-640. doi: 10.1034/j.1600-0587.2002.250510.x.

Elith, J., Kearney, M., and Phillips, S. (2010). The art of modelling range-shifting species. *Methods in Ecology and Evolution* **1**, 330-342. doi: 10.1111/j.2041-210X.2010.00036.x.

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. doi: 10.1146/annurev.ecolsys.110308.120159.

Elith, J., Simpson, J., Hirsch, M., and Burgman, M. A. (2013). Taxonomic uncertainty and decision making for biosecurity: spatial models for myrtle/guava rust. *Australasian Plant Pathology* **42**, 43-51.

FAO, IIASA, ISRIC, ISSCAS, and JRC (2012). Harmonized World Soil Database (version 1.2). (Ed. FAO): Rome, Italy and Iiasa, Laxenburg, Austria.)

Fernández, M., Hamilton, H., Alvarez, O., and Guo, Q. (2012). Does adding multi-scale climatic variability improve our capacity to explain niche transferability in invasive species? *Ecological Modelling* **246**, 60-67. doi: <http://dx.doi.org/10.1016/j.ecolmodel.2012.07.025>.

Fitzpatrick, M. C., Weltzin, J. F., Sanders, N. J., and Dunn, R. R. (2007). The biogeography of prediction error: why does the introduced range of the fire ant over-predict its native range? *Global Ecology and Biogeography* **16**, 24-33. doi: 10.1111/j.1466-822x.2006.00258.x.

Franklin, J. (1995). Predictive vegetation mapping: Geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography* **19**, 474-499. doi: 10.1177/030913339501900403.

Gallant, J. C. and Dowling, T. I. (2003). A multiresolution index of valley bottom flatness for mapping depositional areas. *Water Resources Research* **39**. doi: 10.1029/2002wr001426.

Gallant, J. C. and Wilson, J. P. (2000). Primary Topographic Attributes. In 'Terrain Analysis: Principles and Applications'. (Eds J. P. Wilson and J. C. Gallant) pp. 51-86. (John Wiley and Sons: New York.)

Gevrey, M. and Worner, S. P. (2006). Prediction of global distribution of insect pest species in relation to climate by using an ecological informatics method. *Journal of Economic Entomology* **99**, 979-986.

Glen, M., Alfenas, A. C., Zauza, E. A. V., Wingfield, M. J., and Mohammed, C. (2007). *Puccinia psidii*: a threat to the Australian environment and economy—a review. *Australasian Plant Pathology* **36**, 1-16.

Guisan, A. and Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling* **135**, 147-186. doi: 10.1016/s0304-3800(00)00354-9.

Heersink, D. K., Caley, P., Paini, D. R., and Barry, S. C. (2015). Quantifying the establishment likelihood of invasive alien species introductions through ports with application to honeybees in Australia. *Risk Analysis* **in press**.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* **25**, 1965-1978. doi: 10.1002/joc.1276.

Hijmans, R. J., Phillips, S., Leathwick, J., Elith, J., and Hijmans, M. R. J. (2015). Package 'dismo'.

Hof, A. R., Jansson, R., and Nilsson, C. (2012). The usefulness of elevation as a predictor variable in species distribution modelling. *Ecological Modelling* **246**, 86-90. doi: <http://dx.doi.org/10.1016/j.ecolmodel.2012.07.028>.

Huntley, B., Berry, P. M., Cramer, W., and McDonald, A. P. (1995). Modelling present and potential future ranges of some European higher plants using climate response surfaces. *Journal of Biogeography* **22**, 967-1001. doi: 10.2307/2845830.

Kearney, M., Phillips, B. L., Tracy, C. R., Christian, K. A., Betts, G., and Porter, W. P. (2008). Modelling species distributions without using species distributions: the cane toad in Australia under current and future climates. *Ecography* **31**, 423-434. doi: 10.1111/j.0906-7590.2008.05457.x.

Kearney, M. and Porter, W. (2009). Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters* **12**, 334-350. doi: 10.1111/j.1461-0248.2008.01277.x.

Kearney, M. R., Isaac, A. P., and Porter, W. P. (2014a). microclim: Global estimates of hourly microclimate based on long-term monthly climate averages. *Scientific data* **1**. doi: 10.1038/sdata.2014.6.

- Kearney, M. R., Shamakh, A., Tingley, R., Karoly, D. J., Hoffmann, A. A., Briggs, P. R., and Porter, W. P. (2014b). Microclimate modelling at macro scales: a test of a general microclimate model integrated with gridded continental-scale soil and weather data. *Methods in Ecology and Evolution* **5**, 273-286. doi: 10.1111/2041-210X.12148.
- Kriticos, D. J., Morin, L., Leriche, A., Anderson, R. C., and Caley, P. (2013). Combining a climatic niche model of an invasive fungus with its host species distributions to identify risks to natural assets: *Puccinia psidii* Ssensu Lato in Australia. *PLoS ONE* **8**, e64479. doi: 10.1371/journal.pone.0064479.
- Kriticos, D. J., Webber, B. L., Leriche, A., Ota, N., Macadam, I., Bathols, J., and Scott, J. K. (2012). CliMond: global high-resolution historical and future scenario climate surfaces for bioclimatic modelling. *Methods in Ecology and Evolution* **3**, 53-64. doi: 10.1111/j.2041-210X.2011.00134.x.
- Kriticos, D. J., Yonow, T., and McFadyen, R. E. (2005). The potential distribution of *Chromolaena odorata* (Siam weed) in relation to climate. *Weed Research* **45**, 246-254. doi: 10.1111/j.1365-3180.2005.00458.x.
- Landesman, W. J., Nelson, D. M., and Fitzpatrick, M. C. (2014). Soil properties and tree species drive β -diversity of soil bacterial communities. *Soil Biology and Biochemistry* **76**, 201-209. doi: <http://dx.doi.org/10.1016/j.soilbio.2014.05.025>.
- Lawson, C. R., Bennie, J., Hodgson, J. A., Thomas, C. D., and Wilson, R. J. (2014). Topographic microclimates drive microhabitat associations at the range margin of a butterfly. *Ecography* **37**, 732-740. doi: 10.1111/ecog.00535.
- Longmore, R., Busby, J. R. J. R., and Fauna, A. B. o. F. a. (1986) 'Atlas of elapid snakes of Australia.' 1st edn. (Australian Government Publishing Service.)
- Low-Choy, S., O'Leary, R., and Mengersen, K. (2009). Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models. *Ecology* **90**, 265-277.
- Magarey, R. D., Fowler, G. A., Borchert, D. M., Sutton, T. B., Colunga-Garcia, M., and Simpson, J. A. (2007). NAPFAST: an internet system for the weather-based mapping of plant pathogens. *Plant Disease* **91**, 336-345.
- Matsuki, M., Kay, M., Serin, J., Floyd, R., and Scott, J. K. (2001). Potential risk of accidental introduction of Asian gypsy moth (*Lymantria dispar*) to Australasia: effects of climatic conditions and suitability of native plants. *Agricultural and Forest Entomology* **3**, 305-320. doi: 10.1046/j.1461-9555.2001.00119.x.
- McBride, M. F., Fidler, F., and Burgman, M. A. (2012). Evaluating the accuracy and calibration of expert predictions under uncertainty: predicting the outcomes of ecological research. *Diversity and Distributions* **18**, 782-794. doi: 10.1111/j.1472-4642.2012.00884.x.
- McInerny, G. J. and Purves, D. W. (2011). Fine-scale environmental variation in species distribution modelling: regression dilution, latent variables and neighbourly advice. *Methods in Ecology and Evolution* **2**, 248-257. doi: 10.1111/j.2041-210X.2010.00077.x.

- McKenzie, D., Peterson, D. W., Peterson, D. L., and Thornton, P. E. (2003). Climatic and biophysical controls on conifer species distributions in mountain forests of Washington State, USA. *Journal of Biogeography* **30**, 1093-1108.
- Mellert, K. H., Fensterer, V., Küchenhoff, H., Reger, B., Kölling, C., Klemmt, H. J., and Ewald, J. (2011a). Hypothesis-driven species distribution models for tree species in the Bavarian Alps. *Journal of Vegetation Science*, no-no. doi: 10.1111/j.1654-1103.2011.01274.x.
- Mellert, K. H., Fensterer, V., Kuechenhoff, H., Reger, B., Koelling, C., Klemmt, H. J., and Ewald, J. (2011b). Hypothesis-driven species distribution models for tree species in the Bavarian Alps. *Journal of Vegetation Science* **22**, 635-646. doi: 10.1111/j.1654-1103.2011.01274.x.
- Mesgaran, M. B., Cousens, R. D., and Webber, B. L. (2014). Here be dragons: a tool for quantifying novelty due to covariate range and correlation change when projecting species distribution models. *Diversity and Distributions* **20**, 1147-1159. doi: 10.1111/ddi.12209.
- Morrison, L. W., Porter, S. D., Daniels, E., and Korzukhin, M. D. (2004). Potential global range expansion of the invasive fire ant, *Solenopsis invicta*. *Biological Invasions* **6**, 183-191.
- New, M., Lister, D., Hulme, M., and Makin, I. (2002). A high-resolution data set of surface climate over global land areas. *Climate Research* **21**, 1-25. doi: 10.3354/cr021001.
- Nyström Sandman, A., Wikström, S. A., Blomqvist, M., Kautsky, H., and Isaeus, M. (2013). Scale-dependent influence of environmental variables on species distribution: a case study on five coastal benthic species in the Baltic Sea. *Ecography* **36**, 354-363. doi: 10.1111/j.1600-0587.2012.07053.x.
- Olwoch, J. M., Rautenbach, C. J. D., Erasmus, B. F. N., Engelbrecht, F. A., and van Jaarsveld, A. S. (2003). Simulating tick distributions over sub-Saharan Africa: the use of observed and simulated climate surfaces. *Journal of Biogeography* **30**, 1221-1232.
- Owens, H. L., Campbell, L. P., Dornak, L. L., Saupe, E. E., Barve, N., Soberón, J., Ingenloff, K., Lira-Noriega, A., Hensz, C. M., and Myers, C. E. (2013). Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas. *Ecological Modelling* **263**, 10-18. doi: 10.1016/j.ecolmodel.2013.04.011
- Parmesan, C., Root, T. L., and Willig, M. R. (2000). Impacts of extreme weather and climate on terrestrial biota. *Bulletin of the American Meteorological Society* **81**, 443-450. doi: 10.1175/1520-0477(2000)081<0443:ioewac>2.3.co;2.
- Pearce, J. L., Cherry, K., Drielsma, M., Ferrier, S., and Whish, G. (2001). Incorporating expert knowledge and fine-scale vegetation mapping into statistical modelling of faunal distribution. *JOURNAL OF APPLIED ECOLOGY* **38**, 412-424.
- Pearman, P. B., Randin, C. F., Broennimann, O., Vittoz, P., van der Knaap, W. O., Engler, R., Le Lay, G., Zimmermann, N. E., and Guisan, A. (2008). Prediction of plant species distributions across six millennia. *Ecology Letters* **11**, 357-369. doi: 10.1111/j.1461-0248.2007.01150.x.

- Pearson, R. G. and Dawson, T. P. (2003). Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography* **12**, 361-371. doi: 10.1046/j.1466-822X.2003.00042.x.
- Peterson, A. T., Williams, R., and Chen, G. (2007). Modeled global invasive potential of Asian gypsy moths, *Lymantria dispar*. *Entomologia Experimentalis et Applicata* **125**, 39-44. doi: 10.1111/j.1570-7458.2007.00603.x.
- Piedallu, C., Gegout, J.-C., Perez, V., and Lebourgeois, F. (2013). Soil water balance performs better than climatic water variables in tree species distribution modelling. *Global Ecology and Biogeography* **22**, 470-482. doi: 10.1111/geb.12012.
- Pitt, J. P. W., Régnière, J., and Worner, S. (2007). Risk assessment of the gypsy moth, *Lymantria dispar* (L), in New Zealand based on phenology modelling. *International Journal of Biometeorology* **51**, 295-305. doi: 10.1007/s00484-006-0066-3.
- Pliscoff, P., Luebert, F., Hilger, H. H., and Guisan, A. (2014). Effects of alternative sets of climatic predictors on species distribution models and associated estimates of extinction risk: A test with plants in an arid environment. *Ecological Modelling* **288**, 166-177. doi: <http://dx.doi.org/10.1016/j.ecolmodel.2014.06.003>.
- Queensland Government (2015). Oriental fruit fly. (Queensland Government Department of Agriculture and Fisheries.)
- Régnière, J., Nealis, V., and Porter, K. (2009). Climate suitability and management of the gypsy moth invasion into Canada. In 'Ecological Impacts of Non-Native Invertebrates and Fungi on Terrestrial Ecosystems'. (Eds D. Langor and J. Sweeney) pp. 135-148. (Springer Netherlands.)
- Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., and Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution* **6**, 366-379.
- Renner, I. W. and Warton, D. I. (2013). Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics* **69**, 274-281. doi: 10.1111/j.1541-0420.2012.01824.x.
- Rödder, D. and Engler, J. O. (2012). Disentangling interpolation and extrapolation uncertainties in species distribution models: a novel visualization technique for the spatial variation of predictor variable collinearity. *2012* **8**.
- Rödder, D., Schmidtlein, S., Veith, M., and Lötters, S. (2009). Alien invasive slider turtle in unpredicted habitat: A matter of niche shift or of predictors studied? . *PLoS ONE* **4**, e7843.
- Schutze, M. K., Aketarawong, N., Amornsak, W., Armstrong, K. F., Augustinos, A. A., Barr, N., Bo, W., Bourtzis, K., Boykin, L. M., CACeres, C., Cameron, S. L., Chapman, T. A., Chinvinijkul, S., Chomič, A., De Meyer, M., Drosopoulou, E., Englezou, A., Ekesi, S., Gariou-Papalexiou, A., Geib, S. M., Hailstones, D., Hasanuzzaman, M., Haymer, D., Hee, A. K. W., Hendrichs, J., Jessup, A., Ji, Q., Khamis, F. M., Krosch, M. N., Leblanc, L. U. C., Mahmood, K., Malacrida, A. R., Mavragani-Tsipidou, P., Mwatawala, M., Nishida, R., Ono, H., Reyes, J., Rubino, D., San Jose, M., Shelly, T. E., Srikachar, S., Tan, K. H., Thanaphum, S., Haq, I., Vijayasegaran, S., Wee, S. L., Yesmin, F., Zacharopoulou, A., and Clarke, A. R. (2015). Synonymization of key pest species within the *Bactrocera dorsalis* species complex (Diptera: Tephritidae): taxonomic changes

based on a review of 20 years of integrative morphological, molecular, cytogenetic, behavioural and chemoecological data. *Systematic Entomology* **40**, 456-471. doi: 10.1111/syen.12113.

Seoane, J., Bustamante, J., and DÍAz-Delgado, R. (2005). Effect of Expert Opinion on the Predictive Ability of Environmental Models of Bird Distribution

Efecto de la Opinión de Experto en la Capacidad Predictiva de Modelos de Distribución de Aves usando Predictores Ambientales. *Conservation Biology* **19**, 512-522. doi: 10.1111/j.1523-1739.2005.00364.x.

Simpson, J. A., Thomas, K., and Grgurinovic, C. A. (2006). Uredinales species pathogenic on species of Myrtaceae. *Australasian Plant Pathology* **35**, 549-562.

Stephens, A. E. A., Kriticos, D. J., and Leriche, A. (2007). The current and future potential geographical distribution of the oriental fruit fly, *Bactrocera dorsalis* (Diptera: Tephritidae). *Bulletin of entomological research* **97**, 369-378.

Storlie, C., Merino-Viteri, A., Phillips, B., VanDerWal, J., Welbergen, J., and Williams, S. (2014). Stepping inside the niche: microclimate data are critical for accurate assessment of species' vulnerability to climate change. *Biology Letters* **10**. doi: 10.1098/rsbl.2014.0576.

Sutherst, R. W. and Maywald, G. F. (1985). A computerized system for matching climates in ecology. *Agriculture Ecosystems & Environment* **13**, 281-299. doi: 10.1016/0167-8809(85)90016-7.

Sutherst, R. W. and Maywald, G. F. (2005). A climate model for the red imported fire ant, *Solenopsis invicta* Buren (Hymenoptera: Formicidae): Implications for invasion of new regions, particularly Oceania. *Environmental Entomology* **34**, 317–335.

Synes, N. W. and Osborne, P. E. (2011). Choice of predictor variables as a source of uncertainty in continental-scale species distribution modelling under climate change. *Global Ecology and Biogeography* **20**, 904-914. doi: 10.1111/j.1466-8238.2010.00635.x.

Tingley, R., Vallinoto, M., Sequeira, F., and Kearney, M. R. (2014). Realized niche shift during a global biological invasion. *Proceedings of the National Academy of Sciences* **111**, 10233-10238. doi: 10.1073/pnas.1405766111.

Tobalske, C. (2002) 'Effects of spatial scale on the predictive ability of habitat models for the green woodpecker in Switzerland.'

Tuanmu, M.-N. and Jetz, W. (2014). A global 1-km consensus land-cover product for biodiversity and ecosystem modelling. *Global Ecology and Biogeography* **23**, 1031-1045. doi: 10.1111/geb.12182.

USDA (2015). Asian Gypsy Moth. United States Department of Agriculture, Animal and Plant Health Inspection Service.

Ward, D. (2009). The potential distribution of the red imported fire ant, *Solenopsis invicta* Buren (Hymenoptera: Formicidae), in New Zealand
New Zealand Entomologist **32**, 67-75.

Williams, K. J., Belbin, L., Austin, M. P., Stein, J. L., and Ferrier, S. (2012). Which environmental variables should I use in my biodiversity model? *International Journal of Geographical Information Science* **26**, 2009-2047. doi: 10.1080/13658816.2012.698015.

Xu, T. and Hutchinson, M. F. (2013). New developments and applications in the ANUCLIM spatial climatic and bioclimatic modelling package. *Environmental Modelling & Software* **40**, 267-279. doi: 10.1016/j.envsoft.2012.10.003.

Zimmermann, N., Edwards, T., Moisen, G., Frescino, T., and Blackard, J. (2007). Remote sensing-based predictors improve distribution models of rare, early successional and broadleaf tree species in Utah. *JOURNAL OF APPLIED ECOLOGY* **44**, 1057-1067.

Zimmermann, N. E., Yoccoz, N. G., Edwards, T. C., Meier, E. S., Thuiller, W., Guisan, A., Schmatz, D. R., and Pearman, P. B. (2009). Climatic extremes improve predictions of spatial patterns of tree species. *Proceedings of the National Academy of Sciences* **106**, 19723-19728.

Zurell, D., Elith, J., and Schröder, B. (2012). Predicting to new environments: tools for visualizing model behaviour and impacts on mapped distributions. *Diversity and Distributions* **18**, 628-634. doi: 10.1111/j.1472-4642.2012.00887.x.

9 Appendices

9.1 Appendix A1: Literature review of papers related to predictor variables and species distributions

	Paper	Species, region, data (bg = background)	Variables (cell size)	Methods (models, testdata..)	Conclusions. Elith comments in [] brackets
1	Aalto <i>et al.</i> (2014). The meso-scale drivers of temperature extremes in high-latitude Fennoscandia				[this is simply on constructing climate variables – i.e. how to model temperature extremes in Fennoscandia (ie 68 to 70°N). Water cover and topography drives min temps whereas elevation drives maxima)
2	Ashcroft <i>et al.</i> (2012a). Combining citizen science, bioclimatic envelope models and observed habitat preferences	Bees (<i>Halictus smaragdulus</i>), native plus Hunter Valley NSW, all species records (46 GBIF plus 1029 Belgian reduced to 688 unique) vs subset (19) based on particular form of bee	WorldClim 19 (2.5arcmin) or common 4 (annual temp & rf, max temp warm, min temp cold) or simple 2 (annual temp and rf).	Models: Maxent defaults used, Testdata: 10fold cv, etc: fitted extra models with Hunter valley data too	[Gives interesting demo of effect of predictors on predicted distribution]. Aim to test a combination of SDMs, citizen science and fine-scale habitat prefs to guide surveys for new occurrences [data might be useful]
3	Ashcroft <i>et al.</i> (2011). An evaluation of environmental factors affecting species	Veg on Illawarra escarpment NSW. 37 canopy species. PA data 600 sites, carefully placed to represent varying	Geology, 10 temp predictors hand-made, fine scale; another 10 from a previous time slice (in case veg affected by past temp)	Models: Maxent (BUT PA data!). Testdata: 30% sites randomly excluded.	Propose methodology: test with evaluation dataset; test across species -> find predictive variables. Use paired t-tests to look at drop in AUC with exclusion of variable, across species. Will be signif even if not across all species. [Some useful but not good to use Maxent]. Geology and winter min temps found imp. "Methods such as

	distributions	veg.			hierarchical partitioning (Mac Nally, 2002) and the ones introduced in this article are designed to increase inference on causal factors rather than identify a single best model for any species."
4	Austin (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches				[this is an overview setting up a template of how to think about models. Section 2.2.3 on "selection of environmental predictors" within the "data model" section is relevant to our thoughts on predictors. Cite Huntley and Prentice et al. in their use of physiologically based variables (mean temperature of the coldest month, annual sum of degree-days above 5 ° C and Priestley-Taylor's (an estimate of the annual ratio of actual to pot'l evapotran) and their rationale for that.
5	Austin & Van Niel (2011b). Improving species distribution models for climate change studies: variable selection and scale	Plants.	Tabulate variables used in 10 examples to show that even if people are using ecological theory to choose predictors the outcomes are quite variable in terms of what they decide to use.	Conceptual model: species abundance ~ f(light, temperature, nutrients, water; CO2;disturbance, biota)	[discusses the fundamental concepts of choosing predictors that are ecophysiologically relevant and gives interesting examples] [Talks about choosing a scale relevant to the processes affecting the species. Make a detailed case for light and its effect on physiological studies]
6	Barbet-Massin & Jetz (2014). A 40-year, continent-wide, multispecies assessment of relevant climate predictors for SDM..	Birds (243 species), USA, BBS = PA data	19 WorldClim plus GDD, PET and MI	<u>Models</u> : Biomod (6 methods), <u>Testdata</u> : spatial (repeated 50% split sample) and temporal evaluations, <u>Etc</u> : first worked out proxy sets then chose 1 variable from within each set, repeating thru all; fit and predicted with the "full" set and minus one. The drop in AUC taken to indicate variable importance.	Temp variables most important (PET, GDD, annual temp); annual precip the most imp precip variable (though MI also useful). Annual predictors more useful than seasonal. Consistent results across spatial and temporal evaluations.
7	Bertelsmeier & Courchamp (2014). Future ant invasions in France	14 ant species, PBG data, global data, considering predictions in France. Presences	6 WorldClim variables, justified by some refs that say that temperature. 10 arcmin (19km) resolution	<u>Models</u> : Maxent for variable selection, ensemble of 5 (1- and 2- class SVMs, neural nets, CART and Maxent) within ModEco platform and used consensus (weighted by AUC). Thresholded	[Modelling not strong – global BG is not defensible, ensemble untested, talked about probabilities even though PBG data, treated output as probabilities in comparing across species, thresholded predictions at 0.5. Some of

		from online and refs, BG 10K random global.		output (p=0.5) for predictions. <u>Testdata</u> : 10-fold cv; AUC.	the references for ants in introduction are interesting.]
8	Besnard <i>et al.</i> (2013). Topographic wetness index predicts the occurrence of bird species ...	Wetland birds (4 passerine species) in Western France. PA data from 64 transects	TWI plus spatial eigenvectors to account for SAC between records	<u>Models</u> : GLMs with only linear fits <u>Testdata</u> : model fit tested with AIC weights for inclusion of TWI or not.	[testing the usefulness of topographic wetness indices (4 versions of them); have entirely missed the Australian literature on this including John Gallant's work] Found TWI potentially useful (though didn't compare against predictors other than different variants of TWI).
9	Bradley <i>et al.</i> (2012). Species detection vs. habitat suitability: Are we biasing habitat suitability models with remotely sensed data?				[Good discussion of the role of remotely sensed data (NDVI, veg indices etc) in species distribution modelling, contrasting its use in plant and animal models, talking about predicting potential vs actual distribution. Interesting paper in general and worth knowing about, but we are unlikely to use these variables in invasive species predictions for potential distributions for exactly the reasons they discuss]
10	Braunisch <i>et al.</i> (2013). Selecting from correlated climate variables: a major source of uncertainty for predicting species distributions under climate change	4 mountain birds Switzerland & part Germany. Data from Swiss Ornith Inst treated as PBG. 10000K random BG	4 climate variables: breeding season and winter temp and rf. Derived from WorldClim; 5 topographic; landcover. Made 5 models varying in which climate variables used – 1 per model for 4, or all for 5 th .	<u>Models</u> : Maxent, GLM, GAM, BRT. <u>Testdata</u> : 10-fold cv AUC plus back-project to 1920's AUC.	Current predictions similar; future diverged. [Adds good discussion to the correlated variables debate; they suggest using the set of climate variables rather than choosing one (though it's not clear whether they think all models can handle that). Demonstrate shortcomings in testing correlations b/n variables in both times without looking at how they actually change].
11	Bucklin <i>et al.</i> (2015). Comparing species distribution models constructed with different subsets of environmental predictors	14 vertebrates (6 birds, 4 mammals, 4 reptiles), Florida, PO data though half-treated as PA by defining range map and asserting all non-P cells were absences tho'	7 predictor sets (4km cells): two with bioclimate (bc) predictors only (8 of the WorldClim variables; one a preselected set and one, an uncorrelated set), and five 'combination' models	<u>Models</u> : Biomod – 5 algorithms – GLM, MARS, GBM, RF and Maxent. (!) <u>Testdata</u> : held out samples (25%, repeated splits), <u>Etc</u> : A prelim step looked at variable importance in models fitted to just predictors within each group (climate, human influence etc) and used this to specify later groups of variables. Models evaluated	[this study has several dubious modelling choices including small number of background points, and it's restricted to just Florida, so limited usefulness]. Found small but consistent improvements in prediction when adding predictors to the "best 4" bc predictors. Human influence predictors improved things [perhaps bias in records an issue?]. [Their maps actually show quite an effect across the diff't predictor

		interchangeably also called them pseudo-absences.	using bc plus additional predictors from: human influence (6), land cover (8 classes, each as a proportion of cell area), extreme weather (8; meteorological events over short 1-7 days times) or noise (8; spatially random data).	on AUC and TSS in held out samples (25%, repeated splits) and by variable importance = correlation b/n fitted values with the real variables vs a permuted one. Also tested similarity of maps (correlation coeff on continuous values and compared with range maps for thresholded maps). Evaluated results via GLMMs	sets, so it's not clear that the result is well analysed.]
12	Chatfield <i>et al.</i> (2010). Combining environmental gradients to explain and predict the structure of demersal fish distributions	Demersal fish, The Recherche Archipelago, southern Western Australia.			[talks about finding important predictors and summarises their results from BRTs for 10 species] Substrate type (reef, sand, cobble) was the most influential variable, and water depth and macroalgal type influenced the probable occurrence of species even over the same substrate type. [interesting as an example of marine but our project is terrestrial so take no further]
13	Dubuis <i>et al.</i> (2013). Improving the prediction of plant species distribution and community composition by adding edaphic to topo-climatic variables	Plants (115 species), Western Alps, Switzerland. 252 veg plots, PA data.	25m cells. Topography (slope, topoposition), climate (degree-days, moisture index of growing season, global solar radiation), plus 7 edaphic: (1) pH; (2) the content of nitrogen and of (3) phosphorus; (4) silt; (5) sand; (6) clay and (7) carbon-to-nitrogen ratio	<u>Models</u> : Biomod using GLM, GAM, BRT, RF – results ensembled. <u>Testdata</u> : split sample (30%) repeated 10 times.	pH, total N found most important. [useful discussion of effect of edaphic variables on plants]
14	Rödder & Engler (2012). Disentangling interpolation and				[This is a paper looking at extrapolation and suggesting a new technique for visualising changes in correlations between variables, based on residuals of linear model fitted to pairs of

	extrapolation accuracies ...				standardised variables within training range.]
15	Fernandez <i>et al.</i> (2012). Does adding multi-scale climatic variability improve our capacity to explain niche transferability in invasive species?	10 invasive species (2 amphibians, 6 plants, 1 bird, 1 insect) worldwide.	Made inter-annual monthly climate variables (std dev, coeff var); also used 19 WorldClim variables	<u>Models</u> : Maxent, default settings; <u>testdata</u> : withheld 50% of data. Probably Maxent's defaults for BG (not stated). AUC.	Looked at improvement in AUC with extra variables. It improved AUC for ~ 60% species [though is this overfitting? – it's a random 50%]. Inter-annual alone not so good as WorldClim alone. [Variables are meant to be available but website doesn't work. Makes sense that variation in climate over years might affect some species, but not sure about the rigour of this particular study]
16	Higgins <i>et al.</i> 2012. A physiological analogy of the niche for projecting the potential distribution of plants	22 European tree species. An interpolated abundance product from forestry plots, reduced to sampled PA.	Water availability (soil) from cgar (useful reference), soil N, WorldClim mean, min, max annual temp. 1km mostly.	<u>Models</u> : Thornley's transport resistance (TTR) model, focussing on carbon and nutrient uptake. Used a genetic algorithm to fit parameters to GIS predictors. <u>Testdata</u> : Fit to data used to fit parameters. AUC, commission errors on thresholded predictions. <u>Etc</u> : can look at limiting factors.	[This is quite different, and an attempt to model distributions via a physiological model. Worth reading. Fitted the parameters using distribution data. Good discussion of issues]. [don't discuss value in projecting to new environments]. [Interesting discussion of the value of a "structurally rigid" model, which to me sounds like CLIMEX's advantage].
17	Hof <i>et al.</i> (2012). The usefulness of elevation as a predictor variable in species distribution modelling	54 mammal 117 plant species. Northern Europe.	WorldClim, elevation, veg classes	<u>Model</u> : Maxent. <u>Testdata</u> : withheld 30%; AUC.	[not a bad intro about why elevation might be something you'd want to use, but from then on lit review and testing is not strong. Found ~ half of 75 studies selected used elevation. In their testing elevation either no effect of elevation or some slight evidence it's better to exclude it, but this seems to largely rely on tests of precision of other published "ranges"; unclear why that's a good test.]
18	Kearney <i>et al.</i> (2014a). microclim: Global estimates of hourly microclimate based on long-term				[this describes the MICROCLIM variables and was released with the dataset. We are basing some of our new variables on these].

	monthly climate averages				
19	Kearney <i>et al.</i> (2014b). Microclimate modelling at macro scales ...				[this describes the calculations behind the MICROCLIM variables in more detail, and presents a test of their accuracy in Australia]
20*	Landesman <i>et al.</i> (2014). Soil properties and tree species drive β -diversity of soil bacterial communities	Soil bacteria, 12 forests in eastern US	Soil properties (pH, moisture, NH ₄ , NO ₃ , organic matter; geographic distance	Model: generalised dissimilarity modelling	Soil pH strongest effect on composition (as estimated via turnover); an effect of the tree species too. [interesting paper]
21	Lawson <i>et al.</i> (2014). Topographic microclimates drive microhabitat associations at the range margin of a butterfly	Butterflies, UK, small study at 16 sites and with many transects	Site-measured: bare ground, host plant cover; site-modelled: 5x5m "solar index" – microclimate model combining topography, radiation balance, windspeed; 5km temperature	Model: generalised linear mixed model (GLMM) in WinBUGs	Shows that fine-scale temperature variation generated by topography drives spatial variation in the microhabitat associations of a thermally constrained butterfly. [An interesting example of the idea of mechanistically derived proximal predictors, though in this case at very fine spatial scale. Very good intro with comparison of the thinking behind correlative vs mechanistic models. Worth reading for that].
22	Low-Choy <i>et al.</i> (2009). Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models				[Interesting paper looking at Bayesian models and including expert opinion. They have an example for species distribution modelling, where they have developed software for showing sites on GIS and electing likely presence-absence or abundance; discuss showing response curves in underlying regression model too].
23	McBride <i>et al.</i> (2012). Evaluating the accuracy and calibration of				[Useful if interested in eliciting expert opinion and methods for doing that]. Use lower and upper bound, best guess and confidence interval. "Experts possess valuable knowledge but may

	expert predictions under uncertainty.				require training to communicate this knowledge accurately. Expert status is a poor guide to good performance. In the absence of training and information on past performance, simple averages of expert responses provide a robust counter to individual variation in performance."
24	McInerny & Purves (2011). Fine-scale environmental variation in species distribution modeling ...	Simulated data			[This is interesting wrt trying to deal with errors in variables, which can include the issue that the scale of your predictors doesn't represent the environment experienced by your species. Use a Bayesian latent variable model (with potential for including multiple species) to deal with the uncertainty. A nice idea, and relevant to the issue that we are likely operating at the wrong scale ecologically. At this stage too difficult to implement perhaps but worth keeping in mind]
25	McKenzie <i>et al.</i> (2003). Climatic and biophysical controls on conifer species distributions in mountain forests of Washington State, USA	Conifer species within Washington State USA.	A range of climate (annual temperature, growing-degree days, annual and seasonal precipitation) to biophysical variables (soil, hydrologic, and solar radiation) derived from climatic variables.	<u>Model</u> : GLM. <u>Testdata</u> : interested in model fit and explanation	[Rather slow and detailed, but an interesting mix of variables and testing of models fitted in different forests asking the question: are the same variables and fitted functions selected across forests? (which they took to imply that more causal variables might have been identified)]. Both climatic and biophysical variables important in most cases; climate first.
26	Mellert <i>et al.</i> (2011a). Hypothesis-driven species distribution models for tree species in the Bavarian Alps	14 tree species, Bavarian Alps,	3 water-related (precop, avail water capacity, waterlogging), 3 energy (temp, radiation), nutrition, geomorphodynamics	<u>Model</u> : GAMs. <u>Testdata</u> : withheld 25% of data. Also tested extensively for model fit (including testing interactions, spatial effects, uneven coverage of gradients), realism of response shapes, and predictions of altitudinal limits.	[A good example of using ecological theory – for choice of predictors started with ecological hypotheses and developed variables based on that. The idea of hypothesis-driven modelling threaded through the paper. Worth a read.]
27	Nyström Sandman <i>et al</i> (2013) . Scale-dependent influence of environmental	5 benthic species (4 macrophytes, one animal) in Swedish Baltic Sea coast. 5 extents (25 to	Salinity, depth, slope, wave exposure and substrate (some site-measured)	<u>Model</u> : GAMs. <u>Testdata</u> : interested in deviance explained and variable importance, not predictive performance.	[This is about scale, but extent, not grain. It looks at the effect of changing extent on the importance of predictor variables]. Conclude that relationship b/n extent and relative importance of predictor variables is complex and depends on

	variables on species distribution..	1500km). 1730 sites subsetting according to extent. 120-200 sites used in each model; PA data.			ecology of species. Indirect variables are likely to become less important at larger spatial extent. Persistent importance of depth across all extents. Make the point that depth is fine grained but still important over wide extents.
28	Olwoch <i>et al.</i> (2003). Simulating tick distributions over sub-Saharan Africa: the use of observed and simulated climate surfaces	4 tick species	Comparing 3 different sources of climate data – 2 interpolated (i.e. technique used for variables in Table 2) and one modelled (60km cell)	<u>Model</u> : a multivariate modelling method of Jeffree modified by Erasmus <u>Testdata</u> : = same as training data	[only interesting as an example of someone comparing climate datasets. The methods for comparing the climate data are interesting. The subsequent modelling of ticks isn't particularly interesting for this project and truth is not known.]
29	Parmesan <i>et al.</i> (2000). Impacts of Extreme Weather and Climate on Terrestrial Biota				[Interesting and gives some ecophysiological data, but focuses on extreme weather which is not something we can deal with in this project. They provide all sorts of interesting evidence re species distributions. Butterflies in Nn hemi: observed northward and upward range shift driven by infrequent and severe climate events impacting populations e.g. by causing breakdown in synchrony of life stages with food hosts. Songbirds in USA; northern limits set by nighttime metabolic requirements. Severe cold snaps can cause death. Migratory birds shift abundance wrt weather. Observed episodic local shifts in population abundances can result in range shifts. Freeze and precip tolerance for plants. Armadillo: rainfall, days below freezing]
30	Pearce <i>et al.</i> (2001). Incorporating expert knowledge and fine-scale vegetation mapping into statistical modelling of faunal				[this looks at whether expert opinion helps in modelling fauna. Mostly no effect found; can be useful in constructing of relevant variables for expressing habitat availability. The difficulty in translating results for this project is that this study had good quality survey data for modelling, cf the usual species data available for

	distribution				biosecurity.]
31	Piedallu <i>et al.</i> (2013). Soil water balance performs better than climatic water variables in tree SDM	37 tree species in France; PA data from 32828 sites	1km grid. Variables well thought out: growing degree days (>5deg), min winter temp, climatic water balance (CWB) = precip – PET)2 variants), 2 variants of soil water balance (SWB); tested 3 seasonal measures	<u>Model</u> : GAMs with 4df on splines. <u>Testdata</u> : seems to be training data. AUC and TSS. Looked at the effect on these of adding EITHER CWB (the indirect variable) or SWB on AUC. Also evaluated mapped predictions	[Very nice study of soil water balance as proximal predictor, cf common substitutes. Good discussion of evidence for it as proximal variable, plus other variables. Interesting comparison across species. Provides good evidence for value of proximal predictors (though it's not tested in a "predict to new environments" context).]
32	Pliscoff <i>et al.</i> (2014). Effects of alternative sets of climatic predictors on species distribution models and associated estimates of extinction risk: A test with plants in an arid environment	Rare plants in Chile and Peru. 13 species of <i>Heliotropium</i> . PO data. 10000 BG points.	Made own climate monthlies @ 1km grid, b/c inaccuracies in WorldClim. Made 6 sets of variables: (1) 19 of 48 monthly; (2) 6 monthly; (3) 19 BIOCLIM (bc); (4) 13 bc; (5) 6 bc; (6) first 6 PCA from #1.	<u>Models</u> : Biomod2 (8 methods) <u>Testdata</u> : 30% split off to test (2 replicate splits). AUC, TSS (also looked at projections into future; won't report here). Analysed TSS and thresholded prediction results with GLMMs. Found effects on spatial patterns of predictions but not on predictive performance.	[good lit review at start re what has been done looking at alternate sets of predictors in SDM. Unfortunately they evaluate thresholded rather than continuous predictions; show that effects do not necessarily show wrt predictive performance at points, but they do give different spatial patterns of predictions.]
33	Rödger <i>et al.</i> (2009). Alien invasive slider turtle in unpredicted habitat: A matter of niche shift or of predictors studied?	Slider turtle. 375 native PO records + 205 invaded range that were successfully reproducing.	1km grid. WorldClim variables. (1) All 19; (2) set of 7 commonly used for other species; (3) subset of 5 based on known physiology of species; also 100 random sets of 7 and 5 to test effect of # variables.	<u>Models</u> : Maxent, Bioclim. <u>Testdata</u> : complicated setup testing on different subsets of invasive records, I think. Also all sorts of analyses of the data, including how the predictions match a "climate envelope" they make for the species.	[This is particularly about invasive species prediction and has logical arguments in introduction about why it's a good idea to target physiologically important variables. The results are a little difficult to interpret because Maxent defaults are used so it's partly a story about overfitting. Still, they show clear effects of choice of predictors on predictions, and they assess the ecophysiological based models as best. They have good physiological info on turtle, which could be used to create other predictors too (i.e. not just WorldClim). Main conclusion is that the

					environmental dataset matters]
34	Seoane <i>et al.</i> (2005). Effect of Expert Opinion on the Predictive Ability of Environmental Models of Bird Distribution	Birds in southern Spain			[more evidence that expert selection of subsets of variables doesn't necessarily improve predictive performance]
35	Storlie <i>et al.</i> (2014). Stepping inside the niche: microclimate data are critical for accurate assessment of species' vulnerability to climate change.	Example is with frogs.			[this is an alternative type of microclimate variable to the one we are proposing in this project. Storlie et al's variable is "micro" in the sense that it is statistically downscaled to a grain and topographical accuracy that represents what the species of interest experiences. Needs microclimate data logged at multiple places, in the types of conditions (eg under logs) experienced by the species]
36	Synes & Osborne. (2011). Choice of predictor variables as a source of uncertainty in continental-scale species distribution modelling under climate change.	Great bustard. 1,453 presence records	30 arc-second (~1km); WorldClim monthly and bioclimatic variables, with new variables made from the monthly ones (e.g. PET, growing degree days). Lots of different variable sets chosen, based on quite extensive reviews of the literature looking at what others have done.	Model: Maxent, default settings. <u>Testdata</u> : 25% split sample, AUC. Predictions thresholded, all with same threshold. Then "map comparison kappa" (kappa comparing maps, allowing for differences in location) calculated.	<p>"Generalized variable sets produce an unmanageable level of uncertainty in species distribution models which cannot be ignored. The use of sound ecological theory and statistical methods to check predictor variables can reduce this uncertainty, but our knowledge of species may be too limited to make more than arbitrary choices."</p> <p>Results: all AUCs high, but higher AUC for models with more variables and smaller predicted areas. Kappa varies substantially across variable sets. Make a good case for importance of comparing maps as well as predictive performance at points.</p> <p>[This is a widely cited paper on choice of 37predictors. Looks at both current and future</p>

					predicted distributions. Useful review under “creation of datasets” of what others have done.
37	Tuanmu & Jetz (2014). A global 1-km consensus land-cover product for biodiversity and ecosystem modeling				[this is the paper referred to in the main text section, about deriving a new landcover product that overcomes some of the shortcomings in the source data]
38	Williams <i>et al.</i> (2012). Which environmental variables should I use in my biodiversity model?				[this is comprehensive in terms of explaining ecophysiological basis for variables. Appears to use expert opinion linked to the ecophysiological theories to categorise their 64 variables into proximal etc. See their Table 2. Advice on how to select a particular set for modelling relies on model selection or testing predictive performance – e.g. “ 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”.
39	Zimmermann <i>et al.</i> (2007). Remote sensing-based predictors improve distribution models of rare, early successional and broadleaf tree species in Utah.				[just including this as a source of information on remote-sensing predictors and their use in SDM. Though compare the discussion with that in Bradley et al.]
40	Zimmermann <i>et al.</i> (2009). Climatic extremes improve				[Interesting paper on use of “extremes”, which in this case is really variability: “ we generated a climate predictor set containing long-term (1961–

	predictions of spatial patterns of tree species.				2006) averages of monthly, seasonal, or annual predictors and standard deviations of the mean values representing extremes”. Nice piece of work and interesting use of variables. Has this been done globally yet? – I think the data might be available in the CRU dataset (there are time series there).]
--	--	--	--	--	--

9.2 Appendix A2: The 35 bioclimatic variables in ANUCLIM

- 1) P1. Annual Mean Temperature
- 2) P2. Mean Diurnal Range (Mean(period max-min))
- 3) P3. Isothermality ($P2/P7$)
- 4) P4. Temperature Seasonality (Coefficient of Variation)
- 5) P5. Max Temperature of Warmest Period
- 6) P6. Min Temperature of Coldest Period
- 7) P7. Temperature Annual Range ($P5-P6$)
- 8) P8. Mean Temperature of Wettest Quarter
- 9) P9. Mean Temperature of Driest Quarter
- 10) P10. Mean Temperature of Warmest Quarter
- 11) P11. Mean Temperature of Coldest Quarter
- 12) P12. Annual Precipitation
- 13) P13. Precipitation of Wettest Period
- 14) P14. Precipitation of Driest Period
- 15) P15. Precipitation Seasonality (Coefficient of Variation)
- 16) P16. Precipitation of Wettest Quarter
- 17) P17. Precipitation of Driest Quarter
- 18) P18. Precipitation of Warmest Quarter
- 19) P19. Precipitation of Coldest Quarter
- 20) P20. Annual Mean Radiation
- 21) P21. Highest Period Radiation
- 22) P22. Lowest Period Radiation
- 23) P23. Radiation Seasonality (Coefficient of Variation)
- 24) P24. Radiation of Wettest Quarter
- 25) P25. Radiation of Driest Quarter
- 26) P26. Radiation of Warmest Quarter
- 27) P27. Radiation of Coldest Quarter
- 28) P28. Annual Mean Moisture Index
- 29) P29. Highest Period Moisture Index
- 30) P30. Lowest Period Moisture Index
- 31) P31. Moisture Index Seasonality (Coefficient of Variation)
- 32) P32. Mean Moisture Index of Highest Quarter MI
- 33) P33. Mean Moisture Index of Lowest Quarter MI
- 34) P34. Mean Moisture Index of Warmest Quarter
- 35) P35. Mean Moisture Index of Coldest Quarter

