Modelling invasive species-landscape interactions using high resolution, spatially explicit models

---

A thesis

submitted in partial fulfilment

of the requirements for the Degree of

Doctor of Philosophy

at Lincoln University

by

Senait Dereje Senay

Lincoln University

2014

Abstract

of a thesis submitted in partial fulfilment of the

requirements for the Degree of Doctor of Philosophy.

Modelling invasive species-landscape interactions using high resolution, spatially explicit models

by

Senait Dereje Senay

Invasive species can cause a wide range of damages from destruction of indigenous and productive ecosystems to introduction of vectors to human and animal diseases. In many countries, measures taken to prevent the establishment of invasive species are known to significantly reduce the potential damage that might be caused. As part of those measures, species distribution models (SDMs) are used to predict suitable habitats for highly invasive species so that appropriate strategies to prevent their establishment and further spread can be designed. When species distribution models (SDMs) are used for practical applications, accounting for their uncertainty becomes a priority. However, despite their wide use, reporting the uncertainty of SDM predictions is not well practiced.

The primary aim of the research in this thesis was to identify and quantify uncertainty associated with model predictions of species distributions. The major research question was, why do different models give dissimilar predictions for the same species and/or location? Discrepancy among model results is one of the major issues that affects the perception of their reliability and their capacity to inform policy decisions. In this thesis, the effect of factors considered to influence model performance and drive uncertainty in model predictions, was investigated. The particular factors were, 1) pseudo absence selection, 2) the individual and combined effect of predictor data, dimension reduction methods, and model types on model performance, and, 3) variation within the occurrence data for a given species. Following these investigations, improved procedures developed in this research were used to, 1) investigate the use of a simple mechanistic model to enhance results of

correlative species distribution models in a hybrid approach and 2) improve a dispersal model that can be used to research the potential spread of an invasive species once it has established in a new habitat.

A multi-factor study to investigate the effect of pseudo-absence selection on model performance showed that not only pseudo-absences affect individual models but also consensus among model predictions. To improve individual model performance as well as model consensus, an improved pseudo-absence selection method was developed that balances the geographic and environmental space for selecting pseudo-absences.

The investigation of the individual and combined effect of predictor data, dimension reduction methods and model types on model performance, showed that the type of model is a major factor that affects model performance. The results of this research showed that the combination of appropriate explanatory variables and dimension reduction could increase individual model performance as well as model consensus. Additionally, novel indices that can be used to assess internal characteristics of the environmental predictors and data-pre-processing methods for optimized model performance, were developed.

Another important factor that contributes to model uncertainty is the reliability of the species occurrence data. While the precision of geographical references used for such data and its effect on model predictions and associated uncertainty, has been well studied, however, variation within the occurrence or presence data for a given species has been less investigated. Two case-studies were used to determine the effects of local adaptation within a species, on model predictions. It was found that apparent local adaptations resulting in ecotypes within a species could affect model predictions. As a result, methods are proposed to detect the effect of within presence data variation and an appropriate method to model potential distributions of species with such variable data, is illustrated.

Following improved procedures proposed in this research, the use of a simple mechanistic model to enhance results from correlative species distribution models was investigated. While a well parameterised mechanistic model for species distribution modelling is the ideal, such models need detailed biological data that are most often not available, especially for many invasive insects. In this study, a simple generalized mechanistic model was used to

complement correlative distribution predictions. The resulting predictions from the hybrid model were shown to facilitate the identification of under- or over-predicted areas by correlative models such that its use resulted in improved overall prediction.

The enhanced protocols developed in this thesis were finally used to improve a dispersal model that can be used to project the spread of an invasive species once it has established in a new habitat by the integration of multiple scale suitability layers to represent a realistic landscape over which the dispersal of a given species can be studied. Selective landscape recoding was used to customize the landscape based on specific species-landscape interactions, to improve dispersal rate estimation and dispersal pattern determination.

This thesis presents novel methods that can be implemented to significantly increase model consensus for species distribution predictions. More important, however, the research highlights the need for implementing multi-model and multi-scenario modelling frameworks to reduce model uncertainty that can result from inappropriate use of modelling components. The findings in this thesis form the basis for research aimed at further improvement of species distribution models to provide more reliable tools for applications in invasive species management, biodiversity protection, environmental sustainability and climate change management.

Keywords: SDMs, invasive species, dispersal, spatial modelling, high resolution, dimension reduction, model uncertainty.

Acknowledgement

First of all, I would like to thank my main supervisor Associate Professor Susan Worner. Thank you Sue for believing in me. You have been an excellent supervisor and a caring mentor. Your passion for science in general and specifically for the use of ecological models in understanding the complex ecological reality is an inspiration that cultivated similar passion in myself. I have one more special privilege I need to thank you for, one I have never taken for granted but exercised it nevertheless, your open door policy. It amazes me how you put us, your students first, no matter how much overwhelmed you are with work. Once again thank you, for sharing your knowledge, and for guiding me though an enjoyable but stressful three years and nine months of my life, I am humbly grateful.

To my supervisory team members: Dr. Michael Rostas, Dr. Stephen Hartley and Dr. Jeff Morisette, and my assessor Professor Philip Hulme I thank you very much for imparting me your valuable advice and comments, especially during my 15th month evaluation, which helped to further refine my work towards a fruitful completion. To Dr. Gabriel Senay, I very much would like to thank you for the useful comments on my PhD proposal; your input has helped me shape my research questions. I also would like to thank you for being my role model growing up. To Dr. Craig Phillips, I am ever grateful that you gave me the opportunity to work with the Great White Butterfly data and for taking the time to review two of my chapters. To Dr. William Monahan, thank you for actively explaining and allowing me to apply your generalized mechanistic model in my proposed hybrid prediction study. I am also thankful for your valuable comments on the work done in Chapter 6.

Dr. Takayoshi Ikeda, I am very much thankful for your patience in walking me through the use of the multi-model framework designed by yourself, Dr. Susan Worner and Dr. Gwenaël Leday. I greatly enjoyed working with it, and I can assure you, your time was not wasted in vain. Dr. Joel Pitt, I am very grateful that you personally showed me how to work with your program MDiG. To Jennifer Pannell, Audrey Lustig and Hossein HA Khandan, I thank you so much for proofreading the manuscript prepared out of the work in Chapter 3. I wish you all the best in your Ph.D. studies.

I have made a lot of friends in New Zealand, and of course it is impossible to name all of you. Your support and friendship means much to me and was vital in my journey of doing this Ph.D. For all colleagues in the Bio-Protection Research Centre thank you for all the good times, the support and the cheering-up that were much needed ingredients for helping me overcome the stress of writing this thesis. I wish you well, my friends. To the Ethiopian community in Christchurch, thank you for being a symbol of home. I wish you all the best, and for the love and integrity of the community to prosper with time, so that it continues to be a comfort for future Ethiopian students who find their way, to do their studies in this beautiful land.

My mother Lt. Col. Yeshihareg Cherinet, my Father, Dereje Senay and my brother Engineer Daniel D. Senay, Thank you for being so loving and supportive. My parents, you are the reason I have taken my studies this far, because you instilled a great love for education in me ever since I was little. I am ever so grateful for it. To my husband and my love Brook Daniel Demissie, no words can express how much you supported me while I was writing this thesis! You are my rock, I truly love you and thank you.

Table of Contents

Abstract ii
Acknowledgement v
Table of Contents vii
Table of Figures x
List of tables xii
1. Introduction 1
1.1 Alien Invasive species 1
1.2 Species Distribution Models (SDMs) and Niche theory 4
1.3 Species distribution modelling approaches 7
1.3.1 Correlative species distribution models 7
1.3.2 Mechanistic process-based species distribution models 11
1.3.3 Hybrid models 13
1.3.4 Complex modelling systems 13
1.4 Data sources for species distribution modelling 14
1.4.1 Primary data 14
1.4.2 Secondary data 15
1.5 Species Dispersal Modelling 17
1.5.1 Species spread and dispersal models 17
1.5.2 Challenges for dispersal modelling 18
1.6 Sources of uncertainty in SDMs 19
1.7 Objectives 21
1.8 Thesis structure 22
2. Review of the invasive species used as case studies 25
2.1 Asian tiger mosquito (*Aedes albopictus*) 26
2.2 Yellow crazy ant (*Anoplolepis gracilipes*) 28
2.3 Western corn rootworm (*Diabrotica virgifera virgifera*) 30
2.4 Pine processionary moth (*Thaumetopoea pityocampa*) 34
2.5 Common yellow jacket wasp (*Vespula vulgaris*) 36
2.6 Great white butterfly (*Pieris brassicae*) 38
3. Novel pseudo-absence selection method for improved species distribution modelling 41
3.1 Introduction 42
3.1.1 Types of pseudo-absence selection methods 43
3.1.2 Proposed area of improvement 44
3.2 Methods 46
3.2.1 Biotic data 46
3.2.2 Environmental data 46
3.2.3 Simple random pseudo-absence selection (SM1) 48
3.2.4 Spatially constrained pseudo-absence points selection (SM2) 49
3.2.5 Environmental pseudo-absences point selection (SM3) 49
3.2.6 Three step pseudo-absence selection method (SM4) 50
3.2.7 Model evaluation and output analysis 53
3.3 Results 56
3.3.1 Pseudo-absences 56
3.3.2 Variable selection 58
3.3.3 Model performance 58
3.3.4 Prediction-reality agreement 59
3.3.5 Sensitivity and Specificity 60

3.3.6 Predicted prevalence, model consensus and habitat suitability 61
3.4 Discussion 64
3.4.1 Variable selection 65
3.4.2 Model performance 66
3.4.3 Model consensus and habitat suitability 67
3.4.4 Implications for future A. albopictus and D. v. virgifera management in New Zealand 68
3.4.5 Does model type matter? 71
3.4.6 Caveats 71
3.5 Summary 72
4. Why do models predict differently for the same species and/or locations? 73
4.1 Introduction 73
4.2 Methods 76
4.2.1. Research design and model conceptualization 77
4.2.2 Predictor data (Abiotic) 78
4.2.3 Dimension reduction 80
4.2.4 Species data (biotic) 84
4.2.5 Relative cover indicators (RCIs) 88
4.2.6 Model types 92
4.2.7 Evaluation and validation 93
4.3 Results 95
4.3.1 Variable selection 95
4.3.2 Multivariate analysis 97
4.3.3 Model components 102
4.3.4 Interactions between model components 104
4.3.5 Relative cover indicators vs model performance 106
4.3.6 Species level model selection 108
4.3.7 Species distribution predictions and their associated uncertainty 109
4.4 Discussion 117
4.4.1 Effects of the major modelling components 117
4.4.2 Prediction evaluation beyond the confusion matrix 122
4.4.3 Prediction uncertainty and model averaging 123
4.4.4 Distribution predictions for the five species in this study 124
4.4.5 Caveats 127
4.5 Conclusion: why models give different predictions for the same species and locations 128
5. Incorporating biological traits and environmental adaptation in correlative species distribution models 131
5.1 Introduction 131
5.1.1 Case study 1: range expansion by D. v. virgifera 135
5.1.2 Case study 2: micro adaptation by P. brassicae 135
5.1.3 Research questions 136
5.2 Methods 136
5.2.1 Predictor dataset, Modelling and Validation (both case studies) 137
5.2.2 Identifying components in occurrence data 138
5.2.3 Merging predictions (both case studies) 145
5.3 Results 146
5.3.1 Testing for significant variation in presence data- case study 1 D. v. virgifera 146
5.3.2 Testing for significant variation in presence data- case study 2 P. brassicae 148
5.3.3 Model performance and muli-model comparisons 152
5.3.4 Combined predictions 153
5.4 Discussion 159
5.4.1 Investigating multi-modality in presence datasets 160
5.4.2 The effect of variable selection 162
5.4.3 Combined predictions 164

5.5 Conclusion 166
6. Hybrid species distribution modelling 169
6.1 Introduction 169
6.1.1 Case study: Prediction of the potential distribution of P. brassicae using a hybrid model 171
6.2 Methods 172
6.2.1 Simple Generalized mechanistic niche model framework 172
6.2.2 Correlative species distribution modelling 180
6.2.3 Hybrid model prediction 182
6.3 Results and discussion 183
6.3.1 Physiological suitability - generalized mechanistic niche model prediction 183
6.3.2 Environmental suitability - correlative species distribution model prediction 188
6.3.3 Hybrid potential species distribution prediction 191
6.4 Summary 201
6.4.1 Considerations about the SDMs used in this study 201
6.4.2 Predictability of a species 202
6.4.3 Setting rules to for hybrid correlative-mechanistic predictions 203
6.4.4 Concluding remarks 204
7. Landscape mapping for spatially explicit species dispersal models 207
7.1 Introduction 207
7.1.1 New Zealand invasion of Pieris brassicae 207
7.1.2 Invasive species-landscape interaction 208
7.2 Methods 211
7.2.1 Study area 211
7.2.2 Biological aspects 212
7.2.3 Spatially explicit dispersal model- MDiG 213
7.2.4 Parameterizing P. brassicae dispersal 215
7.2.5 Building the survival layer 220
7.2.6 Assessing the effect of eradication carried out during the years 2012-2013 228
7.2.7 Simulation 230
7.3 Results 232
7.3.1 Dispersal parameters 232
7.3.2 Comparison of occupancy envelopes 235
7.3.3 Effects of current eradication scheme on dispersal dynamics 243
7.4 Discussion 245
7.4.1 Improved representation of species-landscape interaction 246
7.4.2 Dispersal parameters and model choice 247
7.4.3 The future of P. brassicae in New Zealand 248
7.5 Summary 251
8. General discussion 253
8.1 Uncertainty in species distribution modelling 253
8.1.1 Pseudo-absence data generation (Chapter 3) 253
8.1.2 Data, dimension reduction and model type (Chapter 4) 255
8.1.3 Multi-modality in occurrence data (Chapter 5) 260
8.2 Hybrid Correlative-Mechanistic Modelling (Chapter 6) 263
8.3 Species-landscape interactions (Chapter 7) 264
8.4. Future research 266
8.4.1 Pseudo-absence generation 266
8.4.2 Research design for species distribution models 267
8.4.3 Hybrid modelling 267
8.4.4 Species dispersal modelling 268

8.5 Concluding remarks 269
9. References 271
10. Appendices 288
Appendix 3.1 Pseudo-absence generation for *D. v. virgifera* case study 288
Appendix 3.2 Model consensus maps based on different pseudo-absence methods 289
Appendix 4.1 Automated framework to detect change in variable importance over distance 290
Appendix 4.2 Slope, Aspect and Hillshade data derivation 291
Appendix 4.3 Background on h-NLPCA dimension reduction method 294
Appendix 4.4 Ranks of variables as per proportions of their use in the tested models 296
Appendix 4.5 Comparison of prediction accuracy for different species, dimension reduction, model result combinations 297
Appendix 4.6 Presence and pseudo-absence points in the environmental space 299
Appendix 4.7 Ensemble mean predictions and uncertainty maps 301
Appendix 4.8 Map of uncertainty by modelling components 303
Appendix 4.9 External validation data for *V. vulgaris* in New Zealand 305
Appendix 5.1 Individual component model predictions based on different clusters of presence points. 306
Appendix 6.1 Rescaling the potential niche surface into a physiological suitability surface 308
Appendix 7.1 Data extracted from the Atlas of the Insects of The British Isles 309
Appendix 7.2 Occurrence data accessed from the GBIF database 311
Appendix 7.3 Data sources 312
Appendix 7.4 Reference map of place names in the study area (Chapter 7) 315
Appendix 8.1 Research outputs 316
Appendix 8.2 Author contributions to the manuscripts associated with chapter 3 of this thesis 318

Table of Figures

Figure 2.1: Asian tiger mosquito adult (Aedes albopictus) 27
Figure 2.2: Yellow crazy ant adult on sugar bait (Anoplolepis gracilipes) 29
Figure 2.3: Western corn rootworm adult (Diabrotica v. virgifera) 32
Figure 2.4: Pine processionary moth larvae on a pine tree (Thaumetopoea pityocampa) 35
Figure 2.5: Close up shot of an adult Common yellow jacket wasp (Vespula vulgaris) 37
Figure 2.6: An adult large white butterfly (Pieris brassicae) 39
Figure 3.1: Map of global presence data for A. albopictus and D. v. virgifera 47
Figure 3.2: Variable importance analysis for A. albopictus background data delimitation. 51
Figure 3.3: Boundaries of background datasets extracted from circular buffers drawn at various radii from (A) A. albopictus & (B) D. v. virgifera presence points. 52
Figure 3.4: Pseudo-absence points from the four pseudo-absence selection methods 57
Figure 3.5: Variation on mean AUC values due to model type, pseudo-absence selection method and number and structure of presence data 59
Figure 3.6: Kappa values of models for the four pseudo-absence selection methods and two species dataset. 60
Figure 3.7: The effect of pseudo-absence selection method on mean specificity and sensitivity values. 61
Figure 3.8: Percentages of predicted suitable areas and respective model consensus on predictions in New Zealand. (A), Asian tiger mosquito (A.albopictus) (B), Western corn rootworm (D. v. virgifera) 62
Figure 3.9: Global habitat suitability prediction for Asian tiger mosquito (A. albopictus). 63
Figure 3.10: Global habitat suitability prediction for Western corn rootworm (D. v. virgifera) 63
Figure 3.11: Habitat suitability prediction for Asian tiger mosquito (A. albopictus) in New Zealand. 69
Figure 3.12: Habitat suitability prediction for Western corn rootworm (D. v. virgifera) in New Zealand. 70

Figure 4.1: A hypothetical gradient of potential – realized geographic species distribution outputs aligned with species distribution modelling methods of varying complexity. 74
Figure 4.2: Conceptual model showing factorial research design. 78
Figure 4.3: The network topology of the hierarchical auto-associative neural network with bottleneck architecture used for the h-NLPCA 84
Figure 4.4: Map showing the occurrence data distribution of the five species used in this study. 85
Figure 4.5: Subsets of the global study area with different sets of pseudo-absence points by species, predictor data and dimension reduction method. 87
Figure 4.6: Frequency of selection of individual variables across all models 96
Figure 4.7: Density plot of Kappa, AUC, Sensitivity and Specificity scores for the total 180 models. 98
Figure 4.8: Structure correlations (canonical factor loadings) for the first canonical dimension 99
Figure 4.9 Model mean Kappa scores compared over the four modelling components. Error bars indicate the standard deviation over replicate runs. Bars with different letters within a graph indicate statistically significant differences (Tukey's HSD test, $\alpha = 0.025$ for SP & DR, $\alpha = 0.05$ for P & MT). 101
Figure 4.10 Model mean CV error scores compared by the four different modelling components. Error bars indicate the standard deviation over replicate runs. Bars with different letters within a graph indicate statistically significant differences (Tukey's HSD test, $\alpha = 0.025$ for SP &DR, $\alpha = 0.05$ for P &MT). 101
Figure 4.11: A plot of mean Kappa scores against standard deviation over replicates for species –dimension reduction combinations. 103
Figure 4.12: Variation in model mean Kappa scores according to different Species data (SP), dimension reduction methods (DR) and model type (MT) combinations. Bars with different letters are significantly different (Tukey's HSD test, HSD = 0.45, $\alpha = 0.05$). 104
Figure 4.13: Variation in model mean Kappa scores according to different species data (SP) and model type (MT) combinations. Bars with different letters are significantly different (Tukey's HSD test, HSD = 0.24, $\alpha = 0.05$). 106
Figure 4.14: Model Kappa scores plotted against cross-validation error scores 109
Figure 4.15: Predicted probability of presences for A. albopictus. 110
Figure 4.16: Predicted probability of presences for D. v. virgifera. 111
Figure 4.17: Predicted probability of presences for V. vulgaris. 112
Figure 4.18: Predicted probability of presences for A. gracilipes. 112
Figure 4.19: Predicted probability of presences for T. pityocampa. 113
Figure 4.20 Relative locations of D. v. virgifera presences and the three types of pseudo-absences in the environmental spaces of the three predictor datasets. 114
Figure 4.21: (A) Mean predicted presence across all scenarios for A. albopictus; (B) the associated uncertainty around the mean prediction 116
Figure 4.22: Spatial pattern of variability according to (A) Predictor data (P) (B) Dimension reduction (DR), (C) Model type (MT) and (D) the probability density function of the predicted presences by the three modelling components for A. albopictus. 117
Figure 5.1 An illustration showing the possible effects of variation in environmental ranges obtained from presence data. 134
Figure 5.2 Classification of P. brassicae presence points. 142
Figure 5.3 D .v. virgifera presence dataset: (A) histogram plot of the presence dataset with a fitted normal probability density distribution line, (B) the presence dataset after K-means clustering, and (C) the geographic projection of the clustered presence points. 147
Figure 5.4 the cluster means I and N and standard deviation SI and SN fitted to a mixed random probability density curve (blue line). 148
Figure 5.5: Comparison of the relative positions of aestivating and non-aestivating presence points in feature space of four different bioclimatic variables combinations. 149
Figure 5.6: Distribution directional 1SD and 2SD standard deviational ellipses (SDE) derived from the centre means of aestivating, non-aestivating and combined presences of P. brassicae on the feature space of variables selected according to (A) non-aestivating presences (B) aestivating presences. Green stars show P. brassicae locations in New Zealand. 150
Figure 5.7: Performance of models trained based on presences from the native (A), invaded (B) and unclassed presences (all presences) (C) of D. v. virgifera populations 152
Figure 5.9: Accuracy and Sensitivity scores for the combined prediction for both species using the test data from the component predictions (A) D. v. virgifera (B) P. brassicae. 154
Figure 5.10: Global potential D. v. virgifera distribution (A) direct prediction (B) combined prediction 155
Figure 5.11: Global potential P. brassicae distribution (A) direct prediction (B) combined prediction. 156
Figure 5.12: Potential P. brassicae distribution prediction for New Zealand (A) direct prediction (B) combined prediction 157
Figure 5.13 Comparison of predicted presences with true presences in the environmental spaces 158
Figure 6.1: Potential species distribution prediction according to correlative, mechanistic and hybrid species distribution modelling methods. 171
Figure 6.2: Global presence of P. brassicae. Data courtesy of GBIF. 175
Figure 6.3: The seasonal thermal potential niche of Pieris brassicae bounded by the upper and lower lethal temperature thresholds and its geographic projection. 177
Figure 6.4: The global thermal potential niche of Pieris brassicae. 178

Figure 6.5: The potential niche (PN) and projected realized distribution (RD') of Pieris brassicae (A) in the thermal feature space (B) in the geographic space....185
Figure 6.6: The New Zealand extent physiological suitability surface based on thermal thresholds of P. brassicae derived from the mechanistic niche model....187
Figure 6.7: Model performance for the four scenarios, (A) Kappa scores, (B) AUC scores, (predictions from $pbPSN_{Naes}$ and $pbPSN_{aes}$ are subsets of the pbPSNbt scenario)....188
Figure 6.8: The global potential P. brassicae distribution for the different training dataset scenarios. 190
Figure 6.9: The potential P. brassicae distribution predicted for New Zealand, according to the four varying training data scenarios. The green circles show P. brassicae presences in New Zealand. 191
Figure 6.10: Percentage of predicted presences that are outside the potential niche (PN) by the different correlative model scenarios 191
Figure 6.11: Areas of over-predicted thermal ranges identified as suitable by the four correlative predictions. 192
Figure 6.12: Comparison between global and New Zealand extent predictions of the correlative, mechanistic and hybrid models 197
Figure 6.13: Global hybrid potential P. brassicae distribution prediction 198
Figure 6.14: Hybrid suitability prediction for P. brassicae in New Zealand. Inset maps show close up view of the area where P. brassicae is currently present. All coloured areas are suitable (predicted value >=0.5). Variation shows level of suitability. 199
Figure 7.1: Study area of the P. brassicae dispersal modelling study 211
Figure 7.2: Local dispersal neighbourhoods, (A) Von Neumann shape with range = 1 (B) Von Neumann shape with range = 2 (C) Moore shape with range = 1 (D) Moore shape with range = 2 214
Figure 7.3: Suitability maps used to build the survival layer:(A) Hybrid climate suitability, (B) Land cover suitability and (C) Accumulated growing degree days 224
Figure 7.4: Comparison of geographic detail for urban areas between the land cover dataset (A) the high resolution SPOT Map® imagery (B) and the classified man-made structure layer (C)....226
Figure 7.5: Survival layers used in the P. brassicae dispersal model. 228
Figure 7.6: Eradication management blocks used by the department of conservation in Nelson city and surrounding areas. 229
Figure 7.7: Virtual eradication buffers specified according to the active surveillance –eradication scheme currently used on the ground in the years (A) 2012 and (B) 2013 231
Figure 7.8: distances between pre-existing and newly occupied P. brassicae sites fitted to the Cauchy probability density function of the user defined parameter values for United Kingdom data (A), and New Zealand data (B). 233
Figure 7.9: Maximum likelihood iterations to optimize distance parameter estimation. Black triangles indicate the initial values at the top and the best values at the bottom. 234
Figure 7.10: Percentage of correctly represented presences/absences by the occupancy envelopes of dispersal model with the first survival layer (A) dispersal model with the second survival layer (B). 235
Figure 7.11: Dispersal maps overlaid with P. brassicae presences from field survey data 237
Figure 7.12: Dispersal maps overlaid with P. brassicae absences from field survey data 238
Figure 7.13: Area covered by the simulation occurrence envelopes for various thresholds. (A) Dispersal using the first survival layer with added geographic detail from satellite images in urban areas, (B) survival layer without the higher resolution geographic detail in urban areas. 240
Figure 7.14: Comparison between the dispersal models using the survival layers tested in this study. 241
Figure 7.15: Dispersal coverage for the year 2015, 2020 and 2025 based on survival layer one (left panel) and survival layer two (Right panel) 242
Figure 7.16: Comparison between the control dispersal model and the eradication model replicating the current eradication scheme. Left axis: logarithm of site counts, Right axis: Area (km^2) 243
Figure 7.17: Dispersal coverage for the year 2015, 2020 and 2025 based on survival layer 1 with no virtual eradication (left panel) and after eradication in the years 2012 and 2013 (Right panel) 244

List of tables

Table 3.1: List of variables selected using 4 pseudo-absence selection methods for the two target species....55
Table 3.2: Model performance indices 56
Table 4.1: Variables included in the three predictor datasets used for in this study. 79
Table 4.2: Number of presence points (available/ spatially unique) and distances used to limit background extent before pseudo-absence selection for the three types of predictor dataset used for the five species in this study. 86
Table 4.3: Model performance indices 95
Table 4.4: MANOVA results: modelling component effects on model performance. 99
Table 4.5: Relative occurance area (ROA) of the five species 106
Table 4.6: eROR and eRAR ratios for the 45 training datasets 107

Table 4.7: Kruskal-Wallis statistic for mean Kappa score values....107
Table 4.8: Best and worst model component combinations for the five species in this study....108
Table 5.1 Variables selected according to the different presence data components....138
Table 5.2. Circularity index of the directional standard deviational ellipses computed for the three types of presence data classes on two types of environmental variable feature spaces....151
Table 6.1: Lethal and optimal thermal thresholds of *Pieris brassicae* used in the study....173
Table 6.2: Variables used in the generalized mechanistic niche model....174
Table 6.3: Definitions of the various thermal niche fractions calculated for *P. brassicae*....179
Table 6.4 Presence data specification for the correlative models....181
Table 6.5 *Pieris brassicae* niche fraction estimates based on lethal temperature thresholds....184
Table 6.6 Global coverage of the potential thermal niche and various important thermal ranges estimated for *Pieris brassicae.* 186
Table 6.7: Variables selected for the four different training datasets....189
Table 6.8: Areas predicted by the four correlative models in comparison with the mechanistic niche characterizations....193
Table 7.1 Land cover re-classification according suitability for *P. brassicae*....222
Table 7.2: scheme used to combine different survival layer components....227
Table 7.3: Geographically refernced *P. brassicae* survey data points available for validation....232
Table 7.4: Dispersal parameter estimates used to calibrate the MDiG model for the *P. brassicae* dispersal simulation....234
Table 7.5: Performance scores for the dispersal model based on two different survival layers....236

Chapter 1

1. Introduction

1.1 Alien Invasive species

Alien Invasive species are non-indigenous species that adversely affect habitats and ecosystems they invade. While this definition is the one most accepted by ecologists (Richardson *et al.*, 2000) there have been different terms used to describe such species, especially by the public and the media and sometimes by ecologists from other sub-disciplines. Other terms used are "exotic", "introduced", "non-indigenous", "pests", for example. Such terms do not always refer to the species ability to cause damage in its new habitat or the fact that it is introduced from outside its home range. This inconsistency has created some confusion in the way such species are classified, warranting a study on its own (cf.Colautti & MacIsaac, 2004). In this thesis alien invasive species refer to those species that are introduced into a new habitat mostly through unnatural pathways for example through direct or indirect human agency, and are found to be causing either economic, ecological or health problems in their introduced range irrespective of their status in their native range. Although some species become invasive in their native range, most of the serious economic, health and ecological effects of an invasive species is caused by alien species that are introduced into a new habitat through either natural or manmade mechanisms (Simberloff, 2011).

Invasive species have been associated with various negative impacts especially concerning health throughout recorded history. Some of the great epidemics that caused humanitarian

and economic crises like the Bubonic plague (*Yersinia pestis*), malaria (*Plasmodium spp.*) and West Nile virus (*Flavivirus spp.*) are known to be spread by vectors of alien invasive species (Lounibos, 2002). Even though Charles Elton introduced invasive biology as a distinct discipline of ecology as early as 1958 (Elton, 1958; Richardson & Pyšek, 2008), the number of publications on invasion biology and invasive species control has only increased significantly in the last decade (Kolar & Lodge, 2001; Kenis *et al.*, 2009).

The ecological impacts of invasive species on indigenous ecosystems are well known where they damage and disrupt ecosystems either by driving native species into extinction (Novacek & Cleland, 2001; Gurevitch & Padilla, 2004), cause hybridization (Pejchar & Mooney, 2009; Vink *et al.*, 2010), or ecosystem domination (Simberloff, 1996), but their additional impact on productive ecosystems explains why invasive species are considered one of the greatest threats to global ecosystems after climate change. For example, for nations ill-equipped to fight the advances of invasive species, residents can face livelihood changes as their major source of sustenance, land, is taken over by aggressive and hard to control invasive weed species (Angassa & Oba, 2008b, 2008a; Abate *et al.*, 2009; Rangi, 2009). Invasive species also exacerbate the already grave situation of food security in developing countries (Admasu, 2008; Steiner, 2010). With respect to their economic impact, various publications report large sums of money spent on invasive species with respect to the cost of detection, control and eradication (Evans, 2003; Waage & Mumford, 2008; Saunders *et al.*, 2013). According to a World Bank report (2007) USD 1.5 trillion per annum is incurred globally either due to losses caused by invasive species or for their control. Pimentel *et al.* (2004), who based their study on environmental and economic costs associated with alien invasive species, concluded that USD120 billion/year is spent in USA alone.

Increased global trade, tourism, transportation networks and global aid networks have been mentioned as major causes of accidental invasive species introductions into new habitats (Hulme, 2009). Prior prediction of a potential dangerous invasive species can be difficult as some species do not exhibit invasive behaviour in their own environment but can become widely invasive when in contact with a new habitat such as the case of many insect species that have become economic pests. At times a species is introduced to a new habitat but becomes invasive at a much later time when either climatic or man-made changes alter the

ecosystem making it more favourable (Dukes & Mooney, 1999; Matthews *et al.*, 2000; Crooks, 2005). Therefore, additions of new alien invasive species to new habitats are difficult to prevent because of limitations in prior profiling of such species and the growing commercial trade, international aid and transportation networks among nations (Pitt, 2008). However, in many countries, measures taken to prevent the establishment of invasive species are known to significantly reduce the potential damage that might be caused. The successful eradication of the painted apple moth (*Orgyia anartoides*) in New Zealand (Suckling *et al.*, 2007) and the screwworm fly (*Cochliomyia hominivorax)* in the United States, Mexico (Wyss, 2000) and Libya (Cunningham *et al.*, 1992) are good examples of such successful eradications.

The prolific nature of the invasive species problem has made multidisciplinary collaborations necessary, where many fields of biology are used to understand specific traits that make these species successful (Tsutsui *et al.*, 2000; Lee, 2002; Frankham, 2005; Lefort *et al.*, 2012). Such biological information and theories are incorporated into ecological models to characterise suitable habitats and dispersal dynamics of these species by combining biotic and environmental information and/or climate scenarios (Andersen *et al.*, 2004; Elith *et al.*, 2006; Morisette *et al.*, 2006; Worner & Gevrey, 2006; Watts & Worner, 2008; Pitt *et al.*, 2009; Buisson *et al.*, 2010; Sutherst, 2014). Using ecological models for alien invasive species management has become an integral part of invasive species studies due to recent advancements in computing and increased availability of data. Understanding how a species behaves in its own habitat can give a certain insight into where in the world it might establish (Worner & Gevrey, 2006; Phillips, 2008; Worner *et al.*, 2010). Conducting large scale studies is now much easier with species distribution models and their prediction ability can facilitate taking climate change into account (Zimmermann *et al.*, 2007; Elith & Leathwick, 2009; Buisson *et al.*, 2010). Results from species distribution modelling have also been used directly to optimise AIS control and eradiation campaigns (Gottwald *et al.*, 2001; Anderson, 2005; Raymond *et al.*, 2011; Ruscoe *et al.*, 2011).

1.2 Species Distribution Models (SDMs) and Niche theory

"Present distribution of any species is the result of its capacity to multiply and spread within the limitations of time, physical and biotic barriers to all the regions in the world to which it is ecologically suited"

- Buchsbaum and Buchsbaum (1957)

The first logical step to study invasive species distribution is to understand and characterise their habitat, as the success of any species always depends on the suitability of host environment and availability of host species (Worner & Gevrey, 2006). Ecological models with appropriate underlying mathematical and statistical principles have been used for many years to abstract the complex environment and to quickly reach scientific conclusions where observational and/or experimental field studies are not possible. In this sense, some aspects of invasion biology study would not have been possible if ecological models did not exist (Jorgensen, 1986).

Ecological modelling is also an important component of invasion ecology that is instrumental to re-integrate the subfield with classic ecological theories through its mathematical, spatial and correlational platform that enables comparison and cross analysis of natural ecological processes like succession with invasion events. Davis *et al.* (2001) argued that, invasion ecology has slowed down considerably over the years due to the unfortunate dissociation of the invasion ecology subfield from similar ecological subjects like succession. It is important to acknowledge effective ecological modelling tools can be used not only to incorporate information from other fields of ecology, but also to provide important conclusions derived from the study of invasion ecology to other fields of ecology. Albeit, at the expense of the host habitat, invasion provides a unique opportunity to ecologists to understand the natural world by providing an unplanned experiments of large spatial and temporal expanse (Sax *et al.*, 2007).

Ultimately, ecological models that characterise species distribution, pattern and spread both in spatial and temporal dimension enable ecologists to see the difference in the processes and resulting impacts between natural and human assisted invasion of species into a new

habitat. More importantly, these species distribution models are needed in applied ecology where invasive species monitoring and eradication is necessary.

A species distribution model is a specialized model that combines different environmental variables to characterise or predict a suitable habitat for a specific species (Franklin, 2010a). Modelling geographical species distribution was made possible due to the assumption that species can only exist in geographical areas where certain environmental requirements are fulfilled (Guisan & Thuiller, 2005). Therefore, understanding the niche of a species and the theory associated with the concept is rather important to construct any species distribution model.

The term niche in the ecological sense was first described by Grinnell (1924), and he defined the niche as the entirety of the environmental requirements that allow a species to persist and reproduce in a given habitat. In this Grinnellian niche concept, niches could be occupied by their respective species or could be vacant in case of species extinction until filled by other species. Later, Elton (1958) modified the niche concept by considering the species niche as the functional role a species plays in its community. The Eltonian niche concept encompassed biotic interaction through addressing the species' place in the trophic system of a given habitat as the definition of the species niche. Perhaps, the clearest statement that summarises Grinnellian and Eltonian niches was given by Eugene Odum in his definition of the ecological niche as quoted below.

"… the ecological niche of an organism depends not only on where it lives but also on what it does (how it transforms energy, behaves, responds to and modifies its physical and biotic environment), and how it is constrained by other species. By analogy, it may be said that the habitat is the organism's "address," and the niche is its "profession," biologically speaking. " (Odum, 1971)

While Grinnell and Elton started the conversation on the concept of species niche (Colwell & Rangel, 2009), it is Hutchinson's definition that facilitated the underlying ecological assumptions for species distribution models (Kearney, 2006). Hutchinson (1957), defines a species niche as a multidimensional environmental feature space enveloped by optimum environmental conditions and resources on a multiple axes representing the environmental variables important to the species.

According to niche theory species can only exist in their respective niche which is delimited or defined by physiological, morphological and/or biochemical tolerances to key environmental variable gradients like temperature, photoperiod and relative humidity and/or availability of resources important for survival (Hutchinson, 1978). Such description of the niche allowed comparison of large number of environmental variables and availability of resources even though they may be disconnected in the geographical space. Once, the species distribution is estimated using the niche analysis of the environmental space, results are often projected to geographical space to give meaningful spatial information on species distribution.

However, it is extremely difficult to fully comprehend the complete niche of a species let alone map it in a geographical space due to the complex nature of biological interactions, dispersal limitations and portions of the niche that are not realized in the current climate space that prevent a species from occupying the full extent of its niche (i.e. be at equilibrium) (Colwell & Rangel, 2009). To avoid confusion, some distinctions about a species niche have been made in previous studies to enable specification of what aspect of the niche is being studied (Soberón, 2007; Monahan, 2009). These are the fundamental, potential and realized niche of a species.

The fundamental niche refers to the environmental hyperspace within which a species is supposed to survive and reproduce regardless of whether the combination of these environmental requirements exist in the accessible physical space or not. The fundamental niche therefore is made up of a set of environmental conditions which are currently apparent in the geographical space or might have occurred/occur in the geographical space in the past/future (Soberón & Peterson, 2005). In reality, the fundamental niche is not limited by the presence or absence of species that compete with the target species (Soberón & Peterson, 2005). According to Monahan (2009) the ability of species to survive in novel environmental conditions that do not exist in the current geographic range of the species, might be an adaptation to past climate and environmental conditions.

The potential niche refers to the portion of the fundamental niche that is realized in the physical space or geographic extent of a given habitat at any given time (Jiménez-Valverde

et al., 2008; Monahan, 2009). The potential niche ideally can be fully occupied by a species that is at equilibrium with its niche. I believe the perfect example for such a niche is the human being, where being at the top of the trophic level combined by superior advantage of modifying the physical environment allowed humans to exclude any biotic competitions and cross any physical barriers.

The realized niche refers to the portion of the potential niche where the species is actually present. The realized niche is a result of species not being able to occupy all areas that are suited to them due to inter-specific competition, physical barriers or lack of co-evolved or adapted dispersal agents (Soberón & Peterson, 2005). For species that occupy all geographic areas that possess the environmental conditions needed for their sustenance and reproduction, the realized niche is equal to the potential niche (Monahan, 2009).

There are different types of species distribution models which use different computational methods to characterise and/or predict a species distribution in space and time based on known or inferred niche. For example, a model could be stochastic or deterministic; continuous or discrete; steady state or dynamic; mechanistic or relational; machine learning or conventional. These models can be broadly classified based on how the species (biotic) information is processed in the model to produce species distribution predictions and what aspect of the species niche the model attempts to characterise (Jiménez-Valverde *et al.*, 2008).

1.3 Species distribution modelling approaches

1.3.1 Correlative species distribution models

Correlative species distribution models or as sometimes described, habitat suitability models, characterise unique interrelationships of environmental predictors at species presence points to map out possible candidate host environments elsewhere. This type of habitat suitability modelling that infers environmental requirements for a species from their current or past locations of occurrence is referred as correlative habitat suitability modelling (Kearney & Porter, 2009).

Predictors which are assumed to be an important environmental and geographical variables that affect the survival a species, are the building blocks of any correlative habitat suitability study (Elith & Leathwick, 2009). There is so far no restricted or defined number or type of

predictors used for habitat suitability studies. Predictor choice in suitability models are mostly dependent on their availability to the researcher and the target species studied. However, underestimating or overestimating the effect of geographical and/or environmental predictors in any habitat suitability model, can introduce significant error in model results (Ruckelshaus *et al.*, 1997). Thus, models which fail to incorporate factors that affect the immediate microclimate might under- or over-predict site suitability for certain species (Rich & Weiss, 1991; Kearney & Porter, 2009).

Correlative models can be broadly classified into three categories based on how environmental predictors and species occurrence data are characterised and utilized in the models. These are: 1) simple presence-only models, 2) enhanced presence-only models, and 3) presence-absence models.

Presence-only models are models that require only presence data to map species distribution or calculate a habitat suitability index for the species under study. These models extract the environmental space contained within the available presence points using various distance or polygon rules to predict suitable areas for species (Barbet-Massin *et al.*, 2012). Such models assume that species distribution is dependent solely on climatic limits (Stockwell & Peters, 1999). I refer to these types of models "simple presence-only" as they only take presence location information as their reference to map species distribution without considering environmental variable interactions encompassed within the range defined by the presence points. The most common is the rectilinear envelope for BIOCLIM (Busby, 1986), its algorithm uses a rule that defines the maximum and minimum values of environmental variables observed at presence points, giving rapid insight into the environment of the target species. However it is quite hard to represent the complex interaction between environmental variables that partially define the niche of a species using such simple rules. Other examples of simple presence-only models are newer versions of BIOCLIM with combined use of climate envelopes (Nix, 1986; Busby *et al.*, 1991), Mahalanobis distance envelopes for BIOCLIM by Farber and Kadmon (2003), disjoint environmental envelop for HABITAT by Walker and Cocks (1991), point-to-point similarity metric for convex hulls in DOMAIN (Carpenter *et al.*, 1993), α-hulls by Burgman and Fox (2003), discontinued convex hulls by Lobo *et al.* (2010).

Enhanced presence-only models use species occurrence data coupled with additional background data on environmental variables and their interactions which are key for understanding the geographical distribution of species. These models give a more accurate species distribution prediction as opposed to simple presence-only models which simply map areas encompassed by presence points in a convex hull or a polygon structure in the variable space. Examples of enhanced presence-only models include, MAXENT (Maximum Entropy) (Phillips *et al.*, 2006), ENFA (Ecological Niche Factor Analysis) (Hirzel *et al.*, 2002), PBL (Presence and Background Learning Algorithm) (Li *et al.*, 2011). Enhanced presence-only models use all the background data as a set of potential areas for the species presence. Thus there is no set of selected points labelled as "where the species is absent". These models use the set of presence points as a reference with which the background data is compared. Each cell in the resulting distribution map is assigned a probability of presence based on how similar its corresponding background cell is to the set of presence points either based on a probability distribution or some kind of distance measure depending on the algorithm of the model used.

In ENFA, Hirzel *et al.* (2002) employed a PCA-like method to analyse the interaction between environmental variables both for the global (background) data and for presence points, to extract information on possible environmental interactions and limits that define the species niche. Similarly, in the PBL, Li *et al.* (2011) trained a neural network both on environmental values at presence points and the whole background data to profile interactions between environmental variables. In MAXENT, information on environmental variable interactions is extracted both from the background data and from the set of presence points. MAXENT identifies the potential statistical distribution that best uniformly fits the background data while being constrained by the parameters of the distribution of the presence data along multiple environmental variables (Phillips *et al.*, 2006; Elith *et al.*, 2011). Such enhanced presence-only models utilise the complex interactions between environmental variables which the simple presence-only models do not.

Presence-only models are generally sensitive to biases in presence data as all information regarding geographical occurrence of the species is mainly drawn from presence points (Phillips *et al.*, 2009). Bias in presence records is mainly associated with surveys limited to

easy access areas like sampling along roads or tracks, which leaves out possible presences from various environments in less accessible areas. The use of bias grids or weights that limit the influence of presence points in the prediction of areas further from them, has decreased the vulnerability of these models to bias in presence data. Another proposed modification to reduce bias is to limit the background data sampled within a certain distance of presence points to introduce similar bias to the bias expected in the presence data (Phillips *et al.*, 2009).

Presence-absence models are models that use explicitly defined absence points along with presence points to predict potential geographical distribution of the target species within a given spatial extent. Here, absence points could be true absence points obtained from field survey records where the species is recurrently surveyed and not found (Hanberry *et al.*, 2012). If true absence records are not available, various methods like random sampling (Wisz & Guisan, 2009), geographically constrained random sampling (Poulos *et al.*, 2012), or methodological environmental profiling of background data (Chefaoui & Lobo, 2008) are used to identify a set of points as possible absence points (pseudo-absence points).

There are a large number of models in this category, some regression based models like generalized linear models (GLM), generalized additive models (GAM) and logistic regression have been well used in ecological modelling (Guisan *et al.*, 2002; Hartley *et al.*, 2006), while other novel machine learning and classification models like artificial neural networks (ANN), support vector machines (SVM), naïve Bayes (NB) and many other similar models have been more recently used for ecological modelling (Elith *et al.*, 2006; Kampichler *et al.*, 2010; Lorena *et al.*, 2011). Bearing in mind that there cannot be a strict classification as some modelling frameworks mix various types of algorithms, these models can be roughly classified as classical statistical models and machine learning. One characteristic presence-absence models have in common is that a set of true or pseudo absence locations are needed to model habitat suitability/species distribution. Presence-absence models are less sensitive to bias in the presence data compared with presence-only models because part of the information needed to model their distribution comes from absence/pseudo-absence data (Elith *et al.*, 2006). However, obtaining or generating bias free absence data is also a

challenge that has been acknowledged in previous studies (Chefaoui & Lobo, 2008; Lobo *et al.*, 2010; Barbet-Massin *et al.*, 2012).

Correlative models have a well-known limitation in predicting the whole potential distribution of species because presence points used in these models do not always cover all environmental conditions in which the modelled species can survive (Kearney & Porter, 2009; Dormann *et al.*, 2012). Yet the greatest criticism of correlative models is, the very weak link of the approach to niche-theory which is supposed to be the theoretical background that ties the practice of predicting species distribution with ecological reality (Sutherst, 2014). This disconnect is apparent when model results are used without considering the assumptions within which they were produced. Hirzel and Le Lay (2008) discussed some guidelines that should be followed to ensure such studies explicitly link research with the appropriate ecological theory.

Despite these limitations, correlative models can be very useful for assessing risks of establishment by species for which we have little or no biological information other than their geographical occurrence. In addition to requiring less biotic data, correlative models enable the development and testing of new hypotheses about events and processes that derive species distribution (Worner, 1991; Gotelli, 2000).

1.3.2 Mechanistic process-based species distribution models

Mechanistic models, also called process-based models use physiological variables and parameters to model species response to certain environmental conditions. Such models if appropriately parameterized, should better estimate the potential distribution of species than correlative models because they are independent of the current geographical distribution of the species (Dormann *et al.*, 2012). Species occurrence data used in mechanistic modelling is usually for validation of results and is not involved in the model building process. Mechanistic modelling requires detailed knowledge of the biological requirements of the species to be modelled (Buckley *et al.*, 2010). In particular, the physiological processes that are most limiting to the species survival should be identified and parameterized to accurately describe a species potential distribution. There are a number of established processes-based models that have been used for species distribution

predictions. These mechanistic models vary in the assumptions around the variables and functions that limit species distribution. Some models for example CLIMEX[1] (Sutherst & Maywald, 1985) use environmental thresholds of species by taking the minimum, maximum and optimum values of factors that affect species survival and reproduction. Such factors include, temperature, soil moisture and photoperiod to map geographical areas within the tolerance limits of the species under study. Other models like PHENOFIT (Chuine & Beaubien, 2001) use phenological assumptions where specific climatic conditions required for each phenological stage of the species is matched to identify areas that could allow for all phenological stages of the species. There are also more specific and individual models where nutrient intake and energy consumption of species are explicitly considered to predict species distribution (Kearney *et al.*, 2008; Kearney & Porter, 2009). Such detailed models that are based on direct physiological requirements of the species have advantages over correlation models especially when it comes to extrapolative predictions. However, these models may not be readily used for screening of multiple invasive species that is required for an effective national and regional biosecurity procedures due to the species specific knowledge that is not available (Kearney & Porter, 2009; Rodda *et al.*, 2011).

The high biotic data requirement of many mechanistic models contributes to the need for development of alternative correlative species models which allow analysis and visualization of the correlation between biotic or abiotic phenomena and species distribution (Buckley *et al.*, 2010). Poor understanding of the interaction among abiotic factors and the physiological trait they are presumed to control in mechanistic functions might cause mechanistic models to give inaccurate predictions. For example, the geographical distribution of a certain species with known thermal limits can be mapped by identifying areas that have temperatures within the tolerance limits of the species throughout the year. However, such model specification could under or over-estimate the potential distribution of the species if topographical or other climatic features like solar radiation that affect temperature are not considered. Clearly, measuring and identifying interactions between abiotic factors is not always straightforward and its complexity may vary depending on spatial attributes like extent and resolution (Dormann *et al.*, 2013).

[1] http://www.hearne.co.nz/Software/CLIMEX

1.3.3 Hybrid models

Hybrid models refer to modelling species distribution using more than one modelling approach. Although combining mechanistic and correlative models is what is usually referred as hybrid modelling, incorporating other approaches like food web models and community models with correlative models is also considered hybrid modelling. There have been a number of recommendations from previous studies for increased applications of hybrid models for species distribution modelling to take advantage of the strength of both modelling techniques (Kearney & Porter, 2009; Monahan, 2009; Elith *et al.*, 2010).

In fact, Dormann *et al.* (2012) in their recent publication cautioned that dichotomizing the correlative and mechanistic modelling approaches places a negative connotation in the whole species distribution modelling practice and suggest that both approaches are best represented on a continuum where modellers can mix and match their different functionalities depending on the available data and expertise (but see Kriticos *et al.* (2013) for an opposite opinion). Because disadvantages as well as advantages could propagate from these different approaches, Dormann *et al.* (2012) warned that the potential for compounding the shortcomings of the different approaches should be closely investigated before hybrid models are implemented. Even though there is some discussion regarding hybrid species distribution models, there are not many applied studies implementing them (cf - Buckley *et al.*, 2010; Kearney *et al.*, 2010).

1.3.4 Complex modelling systems

Complex modelling systems consider a number of components like phenological variations, food web interactions and dispersal pathways along with climatic and environmental species requirements to model more precise species distribution over space and time (Grimm *et al.*, 2005). Such systems, although difficult to parameterize, would provide the optimum approach that enables applied scientists to delineate the realized distribution of a species by simultaneously considering factors that limit species from being at equilibrium with their potential niche. Such models have an advantage over both mechanistic and correlative models in that they could be used to predict realized species distributions. Although to a varying degree, both correlative and mechanistic models attempt to describe the potential distribution of species (unless biotic interactions are considered in the latter).

Because of the complexity of the different biotic and abiotic factors and their interactions, such models are usually simulation based rather than analytical. One example is the Ecopath-Ecosim-Ecospace simulation framework (Christensen & Walters, 2004). As much as these models are very informative, they are too complex to generalize over space and species as well as time consuming and costly (Brown *et al.*, 2001). Optimizing and pre-determining the level of abstraction of species interaction and temporal resolution for such models is recommended to keep the balance between maintaining reality and reducing complexity (Worner, 1991) as well have a strong suite of uncertainty measures (Aydin *et al.*, 2005).

1.4 Data sources for species distribution modelling

1.4.1 Primary data

Field data is one of the most reliable and most used data source for habitat suitability studies. Most environmental variables are collected through field data surveys. Especially weather variables like precipitation and temperature are the most recorded variables over a long period of time from established weather stations. Such variables are relatively easy to record at a field station and are also considered the most important variables limiting species distribution (Elith & Leathwick, 2009). Even most traditional agro-ecological zones (AEZ) were classified solely based on the temperature and precipitation values of a region. Consistent field data is very important both to understand environmental patterns as well as climate change and its effect on species distribution.

Field observation data is especially indispensable for precise presence and absence data which is the most important component of correlative habitat suitability models. Even though there are ways where presences can be indirectly inferred or acquired from other data sources it is important to ensure a high proportion of presence data acquired by field observation. The level of certainty or uncertainty of habitat suitability model results depend on how precisely the habitat requirements of a specific species is characterized through accurate presence locations.

There is a great need for accurate geographically referenced species occurrence locations and the current availability of easily accessible global biodiversity data has made a significant contribution to the improvement of species distribution models. Two of the most used

databases are the GBIF[2] (Global Biodiversity Information Facility) and CABI[3] (Commonwealth Agricultural Bureaux International). For mechanistic models, laboratory experiments, field observations and empirical parameters that estimate the effect of different biotic and abiotic factors on species are by far the most important sources of information.

1.4.2 Secondary data

1.4.2.1 GIS and remotely sensed data

Many publications report a major advance in the study of species distribution models after the emergence of GIS systems (Pereira & Itami, 1991; Elith & Leathwick, 2009). The powerful data analysis suites associated with GIS have made data analysis and modelling easier and more accurate. However, GIS systems also have an important role in providing data. GIS data and its analysis gives insight into the complex spatial interrelationships of geo-environmental variables which were not fully appreciated previously.

A GIS system can be used to clean spatially redundant occurrence points with respect to the resolution of the model. The generation of interpolated surfaces of variables like elevation, slope, and aspect in a GIS environment with minimal effort makes a number of important environmental variables available for integration in habitat suitability models.

Another specialized source of GIS data is remote sensing. The emergence of remote sensing has made even more potentially important geo-environmental variables like radiation, vegetation status and various leaf area indices available for species distribution models. Remote sensing also makes data from extremely remote, hostile and inaccessible areas available to ecologists, making the global geographic extent for SDM studies nearly complete (Kerr & Ostrovsky, 2003; Bradley & Mustard, 2006; Roura-Pascual *et al.*, 2006; Zimmermann *et al.*, 2007; Andrew & Ustin, 2009).

1.4.2.2 Herbaria and Museums

Data sources like herbaria and museums along with archaeological findings are probably the only way past distributions of species (existent and extinct) could be modelled. Understanding past distributions of species is key to construct a long-term temporal pattern

[2] http://www.gbif.org/
[3] http://www.cabi.org/

on species distribution which could answer interesting evolutionary and bio-geographical questions. Additionally, information on past distribution of species could inform future species distribution predictions for species that show an apparent trend in geographical range shift over time. As ambitious as it sounds, data from such sources however have their own drawbacks. The major problem with such data is lack of explicit geographical reference in many cases (Elith & Leathwick, 2007).

1.4.2.3 Computer generated projections

Data-projections are specifically important to represent future species distributions. These are especially important to understand possible effects of climate change on species. Even if the practice is mired with opposing recommendations regarding the capability of models to extrapolate into future climates (Thuiller, 2003; Araújo *et al.*, 2005; Franklin *et al.*, 2013; Sutherst, 2014). Species distribution models have been used to understand possible effects of different scenarios of future climates on biological invasions, and they are probably the best way to understand future trends of biological invasions and their effects on food production, livelihood protection and biodiversity. Therefore, it is likely that the shortcomings of future species distribution modelling like the uncertainty of extrapolating in to novel climate space will attract more research rather than discontinuing the whole practice as there are no apparent alternatives currently.

1.5 Species Dispersal Modelling

"All organisms in nature are where we find them because they have moved there! This is true even for the most sedentary of organisms"

- Begon *et al.* (2006)

Species dispersal is an important mechanism many species use for survival (Kokko & Lopez-Sepulcre, 2010). Dispersal evolves in response to natural selection promoting outbreeding and avoidance of kin competition for resources. Dispersal occurs also by external factors such as dispersal agents which transfer progenies from their local habitat to new areas (Begon *et al.*, 2006). The meta-population as a whole depends on species dispersal for re-colonizing declining or extinct sub populations from donor patches in the landscape (Hanski & Gilpin, 1991).

Species dispersal becomes both ecologically and economically of great concern when species overwhelm the host environment in ways that result in from serious impact by the invading species, such as with human assisted species dispersal (Sharov *et al.*, 1995; BenDor *et al.*, 2006; Robinet *et al.*, 2009). Studying species dispersal mechanisms and the ability to predict when and where an already introduced invasive species will go next in the landscape has great advantage for subsequent monitoring and eradication strategies (Urban *et al.*, 2008).

1.5.1 Species spread and dispersal models

There are many models developed to predict species dispersal in different environments. The earliest of such models is the "Island" spatially implicit approach that assumes individuals from a certain population just join a group of 'dispersers' and redistribute among patches (Begon *et al.*, 2006). These models do not take spatial locations into account nor do they recognize variation in the probability of dispersing among individuals. An example of such a model is Levin's model (Hanski & Gilpin, 1991; Begon *et al.*, 2006). Other models like the reaction-diffusion model (Skellam, 1951) and its derivatives consider space as one homogeneous state where individuals can disperse to occupy all the available space (Higgins *et al.*, 1996; Hastings *et al.*, 2005; Morozov *et al.*, 2008). More recently spatially explicit population models (SEPM) have been used to model species dispersal explicitly over space and time.

Emergence of a high volume of remotely sensed, and GIS data has encouraged the development and high use of SEPMs. SEPMs have varied assumptions and underlying algorithms, but have in common the use of spatial information along with dispersal parameters. The application of SEPM can result in more information with respect to spatial pattern as they can represent the heterogeneous nature of the landscape and its effect on dispersal behaviour over time (Ruckelshaus *et al.*, 1997). The development of SEPMs have made the use of models for invasive species spread, for the purpose of monitoring, control and eradication, a reality (Urban *et al.*, 2008).

All models have their own advantages and disadvantages. For example, the earlier spatially implicit models failed to elucidate the effect of spatial parameters on dispersal, but their simplicity promoted understanding species dispersal mechanisms (Begon *et al.*, 2006). On the other hand, the spatially explicit models that consider spatial heterogeneity when estimating dispersal rate and pattern, require immense processing power, parameterization and data (Higgins *et al.*, 1996; Ruckelshaus *et al.*, 1997; Hastings *et al.*, 2005; Pitt, 2008).

1.5.2 Challenges for dispersal modelling

There are a number of theoretical (Johnson & Gaines, 1990) and technical (Higgins *et al.*, 1996) factors that challenge dispersal models to precisely estimate the rate and spatial pattern of species dispersal over a given landscape. Two of these factors that are relevant for this thesis are discussed below.

1.5.2.1 Dispersal pathways

Many studies and reviews have concluded that invasive species spread is primarily caused by the increasing human intervention in terms of transport, trade and aid networks (Gottwald *et al.*, 2001; Hulme, 2009; McGeoch *et al.*, 2010). However, these pathways are not usually modelled directly in species dispersal modelling (Lippitt *et al.*, 2008). Some species dispersal modellers account for human intervention by including human population density as part of a habitat suitability layer (Robinet *et al.*, 2009). Clearly it is now possible to achieve greater precision by accounting for specific pathways through explicit data layers like roads, railroads, campsites, dams etc.(Barney, 2006b). However, incorporating human activity explicitly requires detailed parameterization and individual investigation of the

effect of each pathway on the target species. Case studies are necessary because the widely accepted effects of particular pathways might not hold for all species. For example Mader *et al.* (1990) discussed that they found roads and railways acted as more of a linear barrier to epigeic arthropods as opposed to the usual assumption that such infrastructures acts as a corridor for dispersal (Lippitt *et al.*, 2008).

1.5.2.2 Spatial extent, scale and configuration

The observed spatial distribution of species is a result of the interplay between species dispersal and adaptation to new environments (Begon *et al.*, 2006). Therefore, species distribution is a rather dynamic phenomenon. A challenge for dispersal modelling is to decide the appropriate extent, scale and complexity of the study (Frost *et al.*, 1988). Inappropriate spatial extent in dispersal studies leads to a failure to capture the complete dynamics of the system in which the species dispersal is studied (Saura & Martinez-Millan, 2001). The scale at which dispersal studies are carried out also affect accuracy of results, as the level of landscape detail (complexity) that can be incorporated in the model depends on the scale (Levin, 1992). Because movement over the landscape and carrying capacity of a given area depends on the connectivity, shape or isolation of different patches over a heterogeneous environment (Saura, 2002), configuration of different forms in the landscape and how they are represented in the dispersal model affects the estimation of rates of dispersal and abundance of individuals. As there is no one optimum spatial extent, configuration type or composition that can be used for all cases, each of these parameters have to be determined depending on the characteristics of the species studied (Higgins *et al.*, 1996). Even then finding the right combination of these factors for a given dispersal model is a subject of ongoing research (Plećaš *et al.*, 2014; Steckel *et al.*, 2014).

1.6 Sources of uncertainty in SDMs

Following the increase in popularity of species distribution models a number of studies critical of the way model results are interpreted have surfaced. Such criticism has shifted the focus from measuring individual model performance to multiple model comparisons as well

as investigation into uncertainty[4] within modelling frameworks. Elith *et al* (2002) gave a very detailed review of uncertainty types in models. Dormann *et al* (2008) reported that model components such as data quality, collinearity, model type and variable selection have different levels of contribution towards model uncertainty and emphasised model type has the biggest effect on species distribution prediction. Buisson *et al* (2010) corroborated Dormann *et al.* (2008)'s result, stressing predictor and species data sampling errors could be a considerable source of uncertainty, and also noted model type and climate change scenario as major source of uncertainty in model results. Moreover, confusion matrix based validation techniques that are usually used to assess performance of species distribution models have also been reported as possible source of uncertainty (McPherson *et al.*, 2004; Jiménez-Valverde *et al.*, 2008; Lobo *et al.*, 2010).

From the ongoing discussion in the literature regarding uncertainty in species distribution models, it is apparent that quantifying uncertainty or at least specifying it is critical to support the credibility of model results (Elith *et al.*, 2002; Thuiller, 2004; Araújo & Guisan, 2006; Elith *et al.*, 2006; Hartley *et al.*, 2006; Pearson *et al.*, 2006; Araújo & New, 2007; Austin, 2007; Dormann *et al.*, 2008; Roura-Pascual *et al.*, 2009; Buisson *et al.*, 2010; Venette *et al.*, 2010; Gritti *et al.*, 2013).

[4] Uncertainty here is different from the uncertainty reports given for individual performance (model error); it represents the array of uncertainties that propagate throughout the modelling process starting from data collection, pre-processing, standardizing, variable selection, modelling, space and time variation and even result interpretation. Refer to Elith et al (2002) for a comprehensive list.

1.7 Objectives

The overall aim of this thesis is to investigate specific methodological gaps in habitat and climate suitability models used to predict current and future distributions of invasive species.

Specific aims:

- ✧ To investigate sources of uncertainty in commonly used methodologies associated with correlative species distribution models to provide improved procedures for more accurate species distribution predictions.
- ✧ Evaluate the use of simple mechanistic models to rectify inaccuracies in species distribution predictions by correlative species distribution models.
- ✧ Study the effect of multi-scale integration of data from global, regional and local sources for the development of high-resolution resource landscapes for use in dispersal and eradication simulation studies.

Specific objectives:

1. Evaluate the effect of pseudo-absence selection methods both on individual model performance as well as model consensus among different presence-absence models.

2. Develop an improved pseudo-absence generation technique that balances both geographical and environmental space for use in presence-absence correlative species distribution models.

3. Determine if a wider range of geo-environmental predictors additional to the commonly used temperature and precipitation predictors improve global and regional insect habitat suitability predictions.

4. Investigate the effect of linear and non-linear dimension reduction methods on species distribution model performance.

5. Investigate the effect of model component interactions on species distribution model performance based on a factorial experiment and selected case studies.

6. Develop species distribution prediction assessment indices that complement confusion matrix based validation methods.

7. Investigate the effect of variation in species presence data on model performance and specify methods to detect and handle significant variation in presence datasets.

8. Evaluate the benefits of a hybrid prediction from a correlative model and a simplified mechanistic model as a suitable framework to facilitate improved correlative species distribution predictions.

9. Investigate recoding heterogeneous landscapes with a varying degree of composition across space to improve dispersal rate and pattern predictions.

1.8 Thesis structure

Chapter 1 gives background to the research topics discussed in this thesis along with their objectives.

Chapter 2 provides a brief literature review of the six invasive insects that were used as case studies in this thesis. Two of these insect species have already established in New Zealand while the other four have not.

Chapter 3 gives background information on the different kinds of pseudo-absence generation techniques that are used currently for use in presence-absence correlative species distribution modelling. Areas for improvement recommended by previous studies are highlighted. A new method that improves pseudo-absence generation for global and regional studies is described. Major caveats and possible future areas of research are discussed (Covers objectives 1 & 2).

Chapter 4 discusses the discrepancies between species distribution model predictions using examples in previous studies and the investigations in the present research. Possible causes for discrepancies among models are reported based on the study of five species. New indices developed outside the confusion matrix that could be used alongside widely used validation methods are described (Covers objectives 3, 4, 5 & 6).

Chapter 5 presents a discussion of the effects of multi-modality in presence datasets on model prediction accuracy. Novel methods adapted to detect possible individual components that need to be separately considered during species distribution modelling are described (Covers objective 7).

Chapter 6 gives background to the merits and demerits of using hybrid models for species distribution modelling. A simple framework that can be used to update correlative species distribution model results with a minimally parameterised mechanistic model is described. A case study that investigates the effect of using such a framework on prediction accuracy is given (Covers objective 8).

Chapter 7 A species dispersal case study is used to illustrate the effect of utilizing multiple levels of detail (composition) in the landscapes to improve dispersal rate and pattern estimation. A procedure for obtaining varying landscape details through remote sensing data is also described. Additionally, the research results and new developments reported in previous chapters were incorporated in a case study (Covers objective 9).

Chapter 8 presents a general discussion of all the results reported in this thesis and their contribution to addressing the research topics and issues outlined in the objectives. Concluding remarks as well as recommendations for areas of future research to improve species distribution models for better prediction, monitoring and management of invasive species are given.

Chapter 2

2. Review of the invasive species used as case studies

Six invasive insect species that have varying global prevalence were selected as case studies to demonstrate the various methods developed according to the objectives of this thesis. The selected species were *A. albopictus, A. gracilipes, D. v. virgifera, T. pityocampa, V. vulgaris* and *P. brassicae*. Available occurrence data on invasive species usually depends on the number of studies and survey efforts dedicated to a species. However, because of resource constraints usually disease vectors followed by agricultural pests get most attention with ecological pests coming last. The impact on resource constraints is also reflected among the six species selected as case studies in this thesis. For example, because of its critical health hazard status, extensive research has been undertaken on *A. albopictus* in areas of insect control, habitat mapping, and dispersal. Naturally, there is a greater chance of finding almost complete presence range information on such intensively studied insect compared with those that do not threaten human health.

A brief introduction regarding the geographical distribution, pest/vector status and the currently used control methods of the six species is given below. The biology of the species is not covered here. Rather, detailed environmental or physiological requirements of the respective species and the relevant biological background (when available) is given in the description of the respective case study within the research chapters where modelling its

potential distribution is addressed. In this chapter, however, references to literature are provided that have extensive review on the biology of each species.

2.1 Asian tiger mosquito (*Aedes albopictus*)

2.1.1 Geographical distribution, native and invaded

Aedes albopictus (Skuse, 1894) (Diptera: Culicidae) is commonly known as Asian tiger mosquito based on its striped appearance. *Aedes albopictus* was initially described from India as "the banded mosquito of Bengal" by Skuse (1895). Detailed biological review of *A. albopictus* is given by Hawley (1988). *Aedes albopictus* is native to south-east Asia, Western Pacific and Indian Ocean islands (Gratz, 2004). However, it has already invaded the Americas, Indo-pacific regions, Australia, Europe, the Middle East and Africa (Roiz *et al.*, 2011). *A. albopictus* has been intercepted at least 12 times in New Zealand (Derraik, 2004). This mosquito has continued to spread within its introduced range especially in North America where it is found to displace another human disease vector mosquito species in the same genus *Aedes aegypti* (Gratz, 2004).

Currently, *A. albopictus* has a wide geographical range as it has adapted to both tropical and temperate regions. *A. albopictus* has diapausing eggs in temperate climates, which enabled the species to survive cold climates in temperate regions (Hanson & Craig Jr, 1995). The major pathway for its introduction is through transportation of the drought resistant eggs along with used tyres, used containers and plant material (Caminade *et al.*, 2012).

2.1.2 Hosts and vector status

A. albopictus have adapted to breeding in artificial containers (Hawley, 1988), resulting in colonization of urban and highly populated areas where female mosquitoes can use humans as hosts for blood meals. This species is not restricted to a human host. It is a zoonotic vector that can transmit pathogens from its host animals to humans (Lambrechts *et al.*, 2010).

A. albopictus is a vector of at least 22 arbovirses and some *Dirofilaria* nematodes which cause diseases in humans and animals (Honório *et al.*, 2003; Gratz, 2004; Medlock *et al.*, 2006). The list of pathogens transmitted by *A. albopictus* include Dengue, Chikungunya, West Nile

viruses , eastern equine encephalitis, yellow fever, La Crosse, Japanese encephalitis, Potosi, Jamestone Canyon, Tensaw, Keystone, *Dirofilaria immitis* and *Dirofilaria repens* (Roiz *et al.*, 2011). Apart from being a vector, the mosquito is also generally considered a nuisance with painful bites that often have haemorrhagic appearance on victims not used to *A. albopictus* toxin (Derraik, 2004). Also its day biting habit (Lambrechts *et al.*, 2010) makes it difficult to avoid in areas where it is introduced.

Image credit 1 The Earth Times ©2012

*Figure 2.1: Asian tiger mosquito adult (*Aedes albopictus*)*[5]

2.1.3 Control

Typical *A. albopictus* management includes mechanical eradication for example using trapping. Chemical methods are used especially to kill *A. albopictus* larvae and biological methods usually involve parasitizing adults. Carvalho *et al.* (2013) have suggested that even a seamlessly integrated eradication system will not work if community participation is not involved because availability of untended containers that could act as breeding sites could undo any progress achieved through various eradication methods. According to recent

[5] http://www.earthtimes.org/nature/climate-change-asian-tiger-mosquito-invasive/1949/

findings genetic control could prove to be a promising control strategy. For example, Labbé *et al.* (2012) reported that they were able to alter female genes to produce flightless phenotypic variant females that could be used to suppress an *A. albopictus* population. Other methods currently being developed include realising sterile transgenic male *A. albopictus* into wild populations and inducing decline of fitness at the incident of infection of *A. albopictus* with pathogens as reviewed by Carvalho *et al.* (2013).

2.2 Yellow crazy ant (*Anoplolepis gracilipes*)

2.2.1 Geographical distribution, native and invaded

Anoplolepis gracilipes (Smith, 1857) (Hymenoptera: Formicidae) commonly known as yellow crazy ant, long-legged ant or Maldive ant is one of several invasive ants known as tramp ants for their dominant, aggressive and highly invasive traits. Detailed biological and ecological account of the *A. gracilipes* is given by Drescher (2011). The native range of *A. gracilipes* is not known, however more than 80% of the known species from the genus *Anoplolepis* are exclusively from continental Africa, and A. *gracilipes* is the only species in the genus to extend its range outside Africa and the Arabian Peninsula (Wetterer, 2005). Because the first specimen was recorded from India most publications refer to tropical Asia as *A. gracilipes* native range (Chong & Lee, 2010). Currently, *A. gracilipes* is established in tropical islands of the Indian and Pacific Oceans, India, southern China, southern islands of Japan, western Mexico, Chile, South Africa and Australia (Csurhes & Hankamer, 2012). According to Wetterer (2005) both arid environments and high elevation limit *A. gracilipes* distribution stating that specimens were rarely recorded from elevations above 1200 m a.s.l. *A. gracilipes* was detected in Auckland, New Zealand in 2002 but was eradicated soon after (Pascoe, 2002; Abbott, 2005).

Image credit 2 Eli Sarnat, Forestry Images

*Figure 2.2: Yellow crazy ant adult on sugar bait (*Anoplolepis gracilipes*)*[6]

2.2.2 Pest status

A. gracilipes is an opportunistic feeder with known preference for carbohydrate-rich plant nectars and honeydew (Csurhes & Hankamer, 2012). However protein enriched food sources are required for brood production, such that *A. gracilipes* can target invertebrate and small vertebrate prey (O'Dowd *et al.*, 1999). Documented cases of *A. gracilipes* attack include bird chicks, lizards and new-born mammals (Abbott *et al.*, 2005). *A. gracilipes* gained attention as ecological pest after the implications of its impact on island ecosystems and island communities was demonstrated on Christmas Island (O'Dowd *et al.*, 2003), Seychelles Islands (Hill *et al.*, 2003) and Tokelau Islands (Lester & Tavite, 2004). On Christmas Island, *A. gracilipes* has been responsible for initiating a series of destructive biotic changes on the native biota by preying on native red crabs resulting in high biomass growth of the forest understorey which is normally controlled by the crabs. Additionally, *A. gracilipes* can form a mutualistic relationship with pest scales (Abbott & Green, 2007), which accumulate honey dew on trees causing sooty mould cover which can lead to a canopy dieback . Such a cascade of destructive events caused by such invasive species has been labelled "invasion

[6] http://www.forestryimages.org/browse/detail.cfm?imgnum=5475599

meltdown" by Simberloff (2006). Similar to other tramp ants (example, Argentine ant) *A. gracilipes* forms super colonies and are known to have low intraspecific aggression in their invaded ranges (Abbott, 2005; Chong & Lee, 2010) which enables them to overtake new habitats.

2.2.3 Control

There has not been a successful eradication of a large *A. gracilipes* invasion (Abbott *et al.*, 2005; Abbott *et al.*, 2014). On Christmas Island toxicant baits have been used successfully to target *A. gracilipes* while avoiding side effects on non-target species by placing food sources away from the baits to misdirect other vulnerable species like the endangered red crabs (Hoffmann & O'Connor, 2004). However a more successful campaign was carried out using aerial bait dropping that included remote and inaccessible areas, albeit with a greater cost (Boland *et al.*, 2011). Chemical sprays are not as successful as toxic baits (Haines & Haines, 1979). Some studies have identified possible symbiont microbes of *A. gracilipes* for development of potential transgenic population control (Sebastien *et al.*, 2012).

2.3 Western corn rootworm (*Diabrotica virgifera virgifera*)

2.3.1 Geographical distribution, native and invaded

Diabrotica virgifera virgifera (LeConte, 1868) (Coleoptera: Chrysomelidae, Galerucinae) commonly known as the western corn rootworm (WCR) is a pest beetle known to cause extreme damage on maize crop plantations in Northern America and Mexico (Onstad *et al.*, 1999). Coats (1986) stated that the pest was probably introduced to the North American continent about 1,000 years ago from its tropical native origin in Central America. North America is now considered a native range and source of the recent *D. v. virgifera* disseminations to Europe (Henmerik *et al.*, 2004; Moeser & Vidal, 2004; Miller *et al.*, 2005; Ciosi *et al.*, 2008)

Geographically, *D. v. virgifera* has now spread and established in 20 European countries (Gray *et al.*, 2009) including the more recent report of *D. v. virgifera* detection in Matveyev, the Kurgan region of Russia (EPPO, 2011). East Africa and Eastern Asia are also flagged as potential regions into which *D. v. virgifera* might spread subsequently according to CLIMEX

model predictions performed by a number of authors (Hummel *et al.*, 2008 - references within). The multiple dissemination paths used by *D. v. virgifera* in its spread through Europe are a major source of concern. According to genetic studies conducted by Ciosi *et al.* (2008) and Miller *et al.* (2005) to reconstruct possible paths of introduction, they found out that there is at least three separate introduction of *D. v. virgifera* from North America, ruling out the suggestion that the species spread through Europe as a result of a single accidental introduction.

A number of first sightings of the species in Europe are located in close proximity to airports. For example, a first sighting in Belgrade, Serbia was near an airport in 1992 (Henmerik *et al.*, 2004; Ciosi *et al.*, 2008; Gray *et al.*, 2009), first detection in Italy was in Venice close to an airport in 1998 (Henmerik *et al.*, 2004), and in Paris close to the Roissy airport (Ciosi *et al.*, 2008). The detection history suggests a major role of transport network systems in the initial transatlantic dissemination of *D. v. virgifera*. The subsequent successful spread of *D. v. virgifera* throughout Central and south-eastern Europe is attributed to *D. v. virgifera's* ability for long distance active flights (Coats *et al.*, 1986) and passive transportation by wind and other natural forces (Onstad *et al.*, 1999; Spencer *et al.*, 1999).

2.3.2 Hosts and pest status

D. v. virgifera is one of the three most economically damaging species from the Chrysomelidae (leaf beetle) family (Branson & Krysan, 1981). *D. v. virgifera* is a univoltine species, the larvae are oligophagous and specialises on feeding on corn roots (*Zea mays* L.) (Branson & Krysan, 1981). However, according to a food conversion efficiency study by Moseser & Vidal (2004) to identify possible alternative host weed species for *D. v. virgifera* in Europe, they were able to establish that the larvae can successfully feed and grow to maturity on *T. Aestivum* (Winter Wheat), *S. Bicolor* (Sorghum) and *C. dactylon* (Durva or Bermuda Grass). *D. v. virgifera* females and eggs have been found in plots of oat stubble, alfalfa, and winter wheat double-cropped with soybeans and wheat plots in Illinois and larvae have been observed on sunflower (*Helianthus annuus*), alfalfa (*Medicago sativa*) fields in South-eastern Europe (Gray *et al.*, 2009), and found to be successfully feeding on *Cucurbita pepo* (oil pumpkin) fields in Slovenia (Hummel *et al.*, 2008). Gray *et al.* (2009) also noted that

it is probable that *D. v. virgifera* populations introduced to Europe are able to survive in the complete absence of maize on some European monocot grass species such as , *Setaria* spp.

Image credit 3 German Federal Ministry of Education and Research ©2012

Figure 2.3: Western corn rootworm adult (Diabrotica v. virgifera)[7]

2.3.3 Control

Crop rotation, which used to be an efficient control for *D. v. virgifera* in North America, is now proving ineffective as a result of key adaptations like ovipositing in soybeans (*Glycine max* L.) fields which is the main rotation crop for maize. Also, diapausing at the egg stage enables hibernation until new maize crops are planted the following year (Levine & Oloumi-Sadeghi, 1996; Onstad *et al.*, 1999; Spencer *et al.*, 1999; Gray *et al.*, 2009). Crop rotation was also originally thought to be a viable solution to control new introductions of *D. v. virgifera* to Europe, especially because of the heterogeneous nature of farming practice which involves smaller size croplands and use of more than two rotation crops for maize in Europe (Dillen *et al.*, 2009). However, studies into the selection pressure that led *D. v. virgifera* to quickly adapt to oviposition on the alternative soybean crops in N. America suggest that landscape heterogeneity can have negative impact in providing other alternative hosts worsening the spread of the pest (Onstad *et al.*, 1999; Onstad *et al.*, 2001). Moreover, re-

[7] http://www.gmo-safety.eu/basic-info/139.pest-conquers-europe.html

infestations from adjacent crop fields is highly probable unless grand scale coordinated eradication strategies are adapted, as female adult *D. v. virgifera* are able to make long distance sustained flights of up to 25-40 Km (Coats *et al.*, 1986), Henmerik *et al.*(2004) has also reviewed various studies reporting similar average flight distances over a year.

Parasitism by nematode species like *Steinernema carpocapsae* have been reported to be an effective biological control measure against the endogeic larvae of *D. v. virgifera* (Nishimatsu & Jackson, 1998). Investigation to test additional European nematodes and their efficacy against *D. v. virgifera* larvae by Toepfer *et al.* (2005) reported that some nematodes could be a viable and effective biological control agents. Experimental studies show that the larval and pupal stage of *D. v. virgifera* are susceptible to infection by parasitizing nematodes but not the egg stage (Jackson & Brooks, 1995).

Genetically engineered corn varieties specifically those crossed with genes of *Bacillus thuringiensis* (Bt), a bacterium that produce insecticidal toxins, have been successfully used against *D. v. virgifera* outbreaks in N. America from 1996 onwards. However, the species is still a serious economic pest in the central Corn Belt of North America (Gray *et al.*, 2009; Gassmann *et al.*, 2011). Reports of possible resistance of *D. v. virgifera* to the transgenic variety of corn began to be published around 2003 (Jaffe, 2003). More recently , the resistance of *D. v. virgifera* to Bt corn has been experimentally confirmed (Gassmann *et al.*, 2014), negating the best control method against *D. v. virgifera*.

2.3.4 Summary

It took less than two decades for *D. v. virgifera* to develop resistance that enables it to overcome corn-soybean rotation in North America (Onstad *et al.*, 2003) and it took less than 10 years to spread successfully throughout Central and Southeastern Europe with a continual spread towards the rest of Europe by means of smaller satellite populations (Toepfer *et al.*, 2006). Early results indicate that the newly introduced *D. v. virgifera* in Europe is showing tendencies to host expansion specifically to *C. pepo* in case of Slovenia (Hummel *et al.*, 2008) and *T. Aestivum, S. Bicolor* and *C. dactylon* in Germany (Moeser & Vidal, 2004). Studies on *D. v. virgifera* in its newly introduced range show that there is a major host expansion and the pest can no longer be categorised as universally dependent on

corn plantations (Hummel *et al.*, 2008). The rapid shift both in host species and geographical extent as well as the new behavioural and genetic adaptations together with multiple spread mechanisms, makes it apparent that *D. v. virgifera*, poses a high risk of invasion to potential new habitats.

2.4 Pine processionary moth (*Thaumetopoea pityocampa*)

2.4.1 Geographical distribution, native and invaded

Thaumetopoea pityocampa (Denis & Schiffermuller, 1775) (Lepidoptera: Thaumetopoeidae) is commonly known as the Pine processionary moth due to its group marching behaviour in its larval stage. The most descriptive information of the life history of T. pityocampa is given by Fabre (2012). *T. pityocampa* naturally occurs in Central Asia, the Middle East, North Africa and southern Europe. In Europe, despite its distribution being originally limited to the Mediterranean region, there are a number of more recent studies showing that *T. pityocampa* is expanding to colder nearby regions both with respect to latitude and altitude (Hódar *et al.*, 2003; Robinet *et al.*, 2007). Robinet *et al.* (2007) suggest that climate change, specifically winter warming, is a major factor for the expansion of *T. pityocampa* distribution northwards in Europe. Outbreaks of *T. pityocampa* outside its native range include Brittany and Strasbourg in France and north of Italy (Battisti *et al.*, 2005; Robinet *et al.*, 2007). *T. pityocampa* adult females only live for one day which makes any geographical range expansion gradual (Battisti *et al.*, 2006). However, transportation of *T. pityocampa* eggs and larvae with tree material could increase the spread and threat to currently favourable but unoccupied areas (Robinet *et al.*, 2012).

2.4.2 Hosts and pest status

T. pityocampa is a known pest of *Pinus spp.* (Amezaga, 1997; Battisti *et al.*, 2006) but could also attack *Cedrus spp.* (Brockerhoff *et al.*, 2006). The larvae stage of *T. pityocampa* are gregarious defoliators (Kanat *et al.*, 2005) that also make trees susceptible to secondary pests like wood borers (Amezaga, 1997). *T. pityocampa* is also a nuisance possessing urticating larvae hairs causing irritation to humans which can escalate to allergic reactions to some susceptible individuals (Battisti *et al.*, 2006). Eggs are laid under the base of pine tree needles, larvae develop in a colony protected in silken tents hidden within the forest canopy (Hódar *et al.*,

2003). The larvae are night feeders and march in a head to tail procession to feed on nearby branches and later in search of locations to pupate, hence the name (Devkota & Schmidt, 1990). *T. pityocampa* is univoltine with the larval stage predominantly appearing during winter (Stastny *et al.*, 2006). However a possible life-cycle shift has been described in an isolated *T. pityocampa* population that have larvae feeding in the summer in Portugal (Pimentel *et al.*, 2006). *T. pityocampa* has manifested a phenology shift in its new high altitude invaded areas and it is also reported that *T. pityocampa* expansion into high altitude and latitudes in the northern hemisphere are not constrained by availability of primary or potential hosts (Battisti *et al.*, 2005). *T. pityocampa* pupae are reported to enter a prolonged diapause up to 7 years, the reason for such haphazard response in the population is not known but, it could facilitate dispersal in hard times by allowing the species to escape unfavourable climate (Battisti *et al.*, 1998).

Image credit 4 John H. Ghent, USDA Forest Service, United States ©2004

*Figure 2.4: Pine processionary moth larvae on a pine tree (*Thaumetopoea pityocampa*)*[8]

2.4.3 Control

Mechanical removal of nests and application of oil-based insecticides into larval nests are methods that have been used to control outbreaks (Avtzis & Avtzis, 1999). However, aerial

[8] http://commons.wikimedia.org/wiki/File:Thaumetopoea_pityocampa_larva02.jpg

applications of synthetic and biological insecticides are usually more effective but costly. Biological insecticides based on *Bacillus thuringiensis* are reported to be effective against the larvae of *T. pityocampa* (Battisti *et al.*, 1998; Salvato *et al.*, 2002). Studies using entomopathogenic fungi as biocontrol agents have also reported promising results (Er *et al.*, 2007 & references within). Naturally occurring pathogens of *T. pityocampa* are reported to be only effective on subsequent generations that appear after major outbreaks occur in southern Europe (Battisti, 1988).

2.5 Common yellow jacket wasp (*Vespula vulgaris*)

2.5.1 Geographical distribution, native and invaded

Vespula vulgaris (Linnaeus, 1758) (Hymenoptera: Vespidae) is commonly known as the common wasp or common yellow jacket wasp. Detailed accounts of the biology and natural history of *V. vulgaris* can be found in Spradbery (1973)'s review. It is a Holarctic species occurring in Eurasia and North America. However a study by Carpenter and Glare (2010) confirmed that the North American species referred as *V. vulgaris* was misidentified and is actually *V. alascensis*. Therefore it is not known if the Palearctic *V. vulgaris* is currently found in N. America. *V. vulgaris* have been introduced to Australia and New Zealand (Thomas *et al.*, 1990; Matthews *et al.*, 2000). There is a report of detection in Argentina but has not been confirmed since the report by Masciocchi *et al.* (2010). *V. vulgaris* was introduced into New Zealand in 1978, it is the dominant wasp species in the honeydew beech forests (Donovan, 1984), where its density is estimated to be extremely high (Thomas *et al.*, 1990).

2.5.2 Hosts and pest status

V. vulgaris is an ecological pest especially in New Zealand where it outcompetes native bird species by reducing honeydew resources in Beech forests (Thomas *et al.*, 1990). In the South Island of New Zealand, *V. vulgaris* have almost displaced another invasive wasp *Vespula germanica* (Clapperton *et al.*, 1994). Additionally, this species preys on other invertebrates decimating their populations with cascading ecological effects on wildlife. Thus, *V. vulgaris* not only competes with native birds for food but also has a lasting damaging effect in the forest ecosystems it invades by total re-structuring of food webs. For example, Gardner-Gee and Beggs (2013) suggested that the Avian Convergence Hypothesis that states avian-

honeydew associations form in honeydew abundant biogeographic areas whenever ants are not available does not hold where abundant invasive Vespula wasps are available because they disrupt bird-honeydew associations and form wasp-honey dew associations. Heavy mortality rate of caterpillars and other invertebrates is also reported by Beggs and Rees (1999). Beggs and Rees (1999) suggested that in cases like the *V. vulgaris* invasion of New Zealand , it is important to understand the threshold of ecosystem damage so that control and eradication only focus on keeping populations to the threshold level to avoid continuously using chemical and biological insecticides that might hurt other non-target biota.

Image credit 5 By Tim Evison ©2009

*Figure 2.5: Close up shot of an adult Common yellow jacket wasp (*Vespula vulgaris*)* [9]

2.5.3 Control

Mechanical control of wasps through destruction of nests is possible however this method is labour intensive and difficult when large populations in a natural environment are

[9] http://commons.wikimedia.org/wiki/File%3AVespula_vulgaris_portrait.jpg CC-BY-SA-2.5

involved (Toft & Harris, 2004). Because of this, chemical poison baits are usually used , where protein based poisons are preferred to minimize the effect on non-target insects (Spurr, 1995). Unfortunately, it is impossible to completely eradicate *V. vulgaris* from a given area in a single campaign because foraging workers and their queen from adjacent areas re-colonize treated areas (Beggs *et al.*, 1998). Because of the need for re-application of insecticides to suppress *V. vulgaris* populations in invaded areas, control interventions are expensive and have side effects on the native biota. Biological control included the use of the wasp parasitoid *Sphecophaga vesparum burra* in Australia (Field & Darby, 1991). *S. vesparum* v*esparum* diapausing cocoons released as a biocontrol of *V. vulgaris* and *V. germanica* in New Zealand have only established in two sites (Beggs *et al.*, 2002). A number of fungi species were also proposed as a viable biocontrol agent (Harris *et al.*, 2000). A new possible biological control isolated from *Brevibacillius laterosporus* was reported to have a confirmed insecticidal effect on *V. vulgaris* (Glare *et al.*, 2014).

2.6 Great white butterfly (*Pieris brassicae*)

2.6.1 Geographical distribution, native and invaded

Pieris brassicae (Linnaeus, 1758) (Lepidoptera: Pieridae) commonly known as great cabbage white, large white butterfly or great white butterfly. There are a number of reviews on the biology of *P. brassicae* because it is one of the most studied insects. Detail biology and natural history of *P. brassicae* is exhaustively covered by Feltwell (1982).

P. brassicae is a Palearctic phytophagous insect that is common throughout Europe, North Africa and Asia (Feltwell, 1982). *P. brassicae* was accidentally introduced to Chile in 1972 (Feltwell, 1978) and later to South Africa (Gardiner, 1995). *P. brassicae* has recently established in Nelson region in New Zealand in 2010 (MAF, 2012; Kean & Phillips, 2013b). The transportation of diapausing pupae along with cargoes has been suggested as the likely pathway for possible inter-continental dispersals (Feltwell, 1982; NAPPO, 2002). Although, a human agency is required to cover such inter-continental distances, once *P. brassicae* is introduced in an area it can spread effectively because adults are capable of undertaking

long distance active flights, *P. brassicae* are also migratory insects (Feltwell, 1982; Spieth & Kaschuba-Holtgrave, 1996; Spieth & Cordes, 2012- & references within).

2.6.2 Hosts and pest status

P. brassicae is a known pest of plants of the *Cruciferae* family. Experimental studies have also confirmed that plants from this family are the primary food preference of *P. brassicae* larvae (Ansari *et al.*, 2012). Alternative host plant families of *P. brassicae* include *Capparaceae, Papilionaceae, Resedaceae, Tropaeolaceae, Chenopodiaceae, Euphorbiaceae, Geraniaceae, Liliaceae, Phytolaccaceae, Polygonaceae* and *Umbelliferae* (Feltwell, 1982). *P. brassicae* is often associated with agricultural areas, gardens and green spaces. The larvae is a voracious feeder that can totally defoliate its host plants. An earlier estimate of damage caused by *P. brassicae* states that the cost could be up to $204m in its native range and $18.5m in its then introduced range (Feltwell, 1978). Considering the current expanded geographical range and the increase of commercial monoculture of *Brassica* spp. the damage is likely to be far greater. In New Zealand alone, commercial Brassica crops are estimated to be worth 80 million NZD (DOC, 2013b) which would be under threat if *P. brassicae* spreads through the country.

Image credit 6 S Sepp ©2007

*Figure 2.6: An adult large white butterfly (*Pieris brassicae*)* [10]

[10] http://en.wikipedia.org/wiki/File:Large_white_spread_wings.jpg

2.6.3 Control

P. brassicae is one of the few insect species on which there are extensive biological/environmental studies. Because *P. brassicae* is a model species that is used to study a number of generic insect responses towards the environment as well as chemical and biological insecticides, there is an unusually high amount of biological information compared with what is available for many other invasive insects.

Accordingly, a wide range of biological and chemical insecticides have been effectively used against all developmental stages of *P. brassicae*. The methods included are biological control with viruses, fungi, predators, parasitoids and chemical insecticides of many formulations. Some chemicals, for example, for example, like DDT[11] are not used anymore. An extensive list of the biological agents used is given by Feltwell (1982). Specifically, the branconid wasp *Apanteles glomeratus* is an effective parasitoid of the larval stage which kept *P. brassicae* numbers down in most areas in the United Kingdom and other regions in Europe (Debarma & Firake, 2013). This parasitoid is also present in New Zealand. The granulosis virus (GV) is a widely used effective control of *P. brassicae* in various parts of the world (Bonnemaison, 1965; Hochberg, 1991). In New Zealand both *Cotesia glomerata* and *Pteromalus puparum*, larval and pupal parasites respectively occur in the wild, the *C. glomerata* reared from collected *P. brassicae* larvae are used to control *P. brassicae* as part of an integrated eradication regime (Phillips *et al.*, 2013).

Transgenic control using host plants that express *Bacillus thuringiensis* (Bt) toxin have been shown to be effective against *P. brassicae* larvae (Kjær *et al.*, 2009). However, care should be taken in implementation of pest control through transgenic crops that express Bt toxins. According to a report by Tabashnik *et al.* (2013) five insect pest species were found to be resistant to the Bt variety crops of Corn and Cotton. Jaffe (2003) warned that planting at least 20% of a cropland as a refugia with non-Bt variety crop is essential to deter the development of resistance or at least prolong resistance by pest insects.

Removing unwanted host species from introduced ranges is an important practice to control *P. brassicae* population along with any eradication strategy (Phillips *et al.*, 2013).

[11] dichlorodiphenyltrichloroethane

Chapter 3

3. Novel pseudo-absence selection method for improved species distribution modelling

The results of this chapter are published as

Senay, S. D., Worner, S. P., & Ikeda, T. (2013). Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modelling. PLoS ONE, 8(8), e71218. doi:10.1371/journal.pone.0071218

Abstract

Pseudo-absence selection for spatial distribution models (SDMs) is the subject of ongoing investigation. Numerous techniques continue to be developed, and reports of their effectiveness vary. Because the quality of presence and absence data is key for acceptable accuracy of correlative SDM predictions, determining an appropriate method to characterise pseudo-absences for SDM's is vital. The main methods that are currently used to generate pseudo-absence points are: 1) randomly generated pseudo-absence locations from background data; 2) pseudo-absence locations generated within a delimited geographical distance from recorded presence points; and 3) pseudo-absence locations selected in areas that are environmentally dissimilar from presence points. There is a need for a method that considers both geographical extent and environmental requirements to produce pseudo-absence points that are spatially and ecologically balanced. We use a novel three-step approach that satisfies both spatial and ecological reasons why the target species is likely to find a particular geolocation unsuitable. Step 1 comprises establishing a geographical extent around species presence points from which pseudo-absence points are selected based on analyses of environmental variable importance at different distances. This step gives an ecologically meaningful explanation to the spatial range of background data, as opposed to using an arbitrary radius. Step 2 determines locations that are environmentally dissimilar to the presence points within the distance specified in step one. Step 3 performs K-means clustering to reduce the number of potential pseudo-absences to the desired set by taking the centroids of clusters in the most environmentally dissimilar class identified in step 2. By considering spatial, ecological and environmental aspects, the three-step method identifies appropriate pseudo-absence points for correlative SDMs. We illustrate this method by predicting the New Zealand potential distribution of the Asian tiger mosquito (*Aedes albopictus*) and the Western corn rootworm (*Diabrotica virgifera virgifera).*

3.1 Introduction

Spatial distribution models (SDMs) have been used to model species distribution for conservation, biological control introductions and, particularly, to predict invasive species establishment and spread (Araújo & Peterson, 2012). Correlative SDMs are popular as the alternatives, mechanistic models are not always achievable due to their requirement of extensive knowledge of the environmental and physiological requirements of the species (Peterson, 2006; Kearney & Porter, 2009).

One of the sources of uncertainty in correlative SDM predictions is the lack of true absence information for accurate species distribution model calibration (Soberón & Peterson, 2005; Wisz & Guisan, 2009). Determining true absences for species distribution prediction is a difficult task. A species could be absent for reasons other than simply because the location is not environmentally suitable (Hirzel *et al.*, 2002; Araújo & Peterson, 2012). Possible scenarios include: 1) the species has not reached the locality due to natural or human barriers, 2) the species has not been detected despite being present, or 3) it is excluded due to competition. Other potential reasons could also be that the species has become locally extinct despite the environment being favourable or temporarily absent due to migratory behaviour.

To compensate for the lacking absence information the following two types of modelling methods are used. 1) Presence-only models where models infer suitable areas for the species by utilizing only occurrence points and their association with the environment. These could be simple or enhanced presence-only models depending on usage of the background environmental data and model algorithm (Description given in section 1.3.1). 2) Presence-absence models that use both presence and absence information to predict habitat suitability and/or species distribution. In case of the latter where real absences are not available, various techniques are used to generate pseudo-absence points.

The choice between the above two approaches is often influenced by the quality and quantity of presence data and research objectives such as whether a potential or realized species distribution prediction of the species is required (Jiménez-Valverde *et al.*, 2008).

The disagreement among studies that have evaluated both types of models (Hirzel *et al.*, 2001; Zaniewski *et al.*, 2002; Elith *et al.*, 2006; Phillips *et al.*, 2006; Elith & Leathwick, 2007;

Poulos *et al.*, 2012; Hastie & Fithian, 2013) shows that each type has merits depending on the modelling context, such as: availability of presence data, characteristics of the predictor data and the modelling expertise available.

Presence-only models work best when there is a reasonable sample of presence information for the target species, preferably with minimal bias (Hirzel *et al.*, 2002; Phillips *et al.*, 2006). If the available presence data is incomplete or uncertain, presence-absence models are thought to produce more robust results. That is because absence and/or pseudo-absence points can minimize over-prediction and extrapolation into unknown areas (Brotons *et al.*, 2004; Elith *et al.*, 2006). It is always better, statistically, to develop a model that predicts based on negatives (in our case absences or zeroes) and positives (presences or ones) than only using positives, provided that the negative data are reliable (Manevitz & Yousef, 2002).

Availability of true absence points is very limited in reality, thus to benefit from the advantages of presence-absence models reliable pseudo-absences are required. A number of studies have proposed different, often contradicting pseudo-absence selection methods (Chefaoui & Lobo, 2008; VanDerWal *et al.*, 2009; Wisz & Guisan, 2009; Lobo *et al.*, 2010; Warton & Shepherd, 2010; Barbet-Massin *et al.*, 2012). Even with contrasting recommendations about pseudo-absence selection methods, these studies agree that the quality of pseudo-absence data directly affects the accuracy of model predictions.

3.1.1 Types of pseudo-absence selection methods

3.1.1.1 Simple random pseudo-absence selection

This method involves taking pseudo-absence points from the background data at random usually excluding known presence points. No prior information about the presence and background data is incorporated to the selection procedure (Lütolf *et al.*, 2006; Wisz & Guisan, 2009). A variation of this method is when available true absence records are included along with the selected random pseudo-absence points (Stockwell & Peters, 1999).

3.1.1.2 Pseudo absence points with limited geographical extent

This method involves selection of pseudo-absence points within (or outside) a certain geographic distance from presence points. Some studies use trial and error where pseudo-absence locations are selected from an area encompassed by varying radii around known

presence points. The ideal distance (radius) is chosen based on model performance results (Hirzel *et al.*, 2001; VanDerWal *et al.*, 2009; Lobo *et al.*, 2010; Barbet-Massin *et al.*, 2012). There are also cases where the radius is chosen arbitrarily or based on expert knowledge about the species (Poulos *et al.*, 2012).

3.1.1.3 Pseudo-absence points based on environmental variables

Models that use this method are often referred to as a two-step-pseudo absence selection method. The method involves prior profiling of environmental data into classes (Zaniewski *et al.*, 2002; Engler *et al.*, 2004; Chefaoui & Lobo, 2008) using niche analysis models such as ENFA, MDE (Lobo *et al.*, 2006), BIOCLIM (Farber & Kadmon, 2003), statistical methods like the Poisson point process method (Warton & Shepherd, 2010), or simply removal of the known environmentally suitable locations from background data before selecting pseudo-absences. Once the least suitable areas are identified by such profiling, pseudo-absence points are selected at random. Many studies report increased accuracy using this approach. Moreover, judging from its repeated use in species distribution modelling studies (Chefaoui & Lobo, 2008; Hengl, 2009; Warton & Shepherd, 2010; Hanberry *et al.*, 2012) it seems this method has become a standard.

3.1.2 Proposed area of improvement

Current pseudo-absence selection methods either optimize for better environmental or spatial discrimination. There is no existing method that provides a balance between these two dimensions. Good discrimination between presence and pseudo-absence points in environmental space alone gives models clear information about the domains in which the species could or could not occur. However, if there is no spatial constraint a model is likely to pick up global or larger scale differences rather than local variations that result in "there-are-no-polar-bears-in-the-Sahara" predictions (Lobo *et al.*, 2010). VanDerWal *et al.* (2009) reported that geographical/spatial extents of background data affected the accuracy of model predictions for 12 species. Furthermore, variable importance varied depending on the size and extent of background data (VanDerWal *et al.*, 2009). This result raises two important questions. Does bounding background data at a certain distance from the presence points before pseudo-absence selection affect prediction accuracy? If so, what distance is appropriate for the species and predictor variables involved?

Barve *et al.* (2011) asserted that the distance used to limit background data affects model training, validation and comparisons, all of which are important modelling components that ensure species distributions are predicted appropriately. Perhaps the most comprehensive framework to identify an optimal background distance was proposed by Barve *et al.* (2011). Their framework involves determining background distance through a dynamic dispersal model that identifies the possible geographically accessible area to a species within a given time (t). As informative and detailed as this method is, it unfortunately cannot be used for all cases of species distribution prediction studies. That is primarily because the above explained model is geared towards determining the realized distribution of a species by specifying areas that are physically accessible to the species. However, areas that are identified as inaccessible according to a time specified dispersal model could be made accessible through human assisted long distance dispersal. According to Lobo *et al.* (2010) decisions about giving either spatial or environmental space more weight while selecting pseudo-absence points depends on whether the objective of the study is to model the realized or potential distribution of the target species.

This study progresses the ideas proposed by Lobo *et al.* (2010) and Barve *et al.* (2011) regarding the need for incorporating geographical constraint for pseudo-absence selection methods, and provides a tested protocol that incorporates the use of geographically and environmentally balanced pseudo-absence points for improved habitat suitability analysis and potential species distribution predictions.

Comparisons are made between model predictions based on the newly proposed method and predictions that used the three commonly used pseudo-absence selection techniques. Presence data of two species (*Aedes albopictus,* and *Diabrotica v. virgifera)* with varying relative occurrence area were used to test the pseudo-absence selection methods both in scenarios where presence points are abundant or scarce in relation to the study area. The comparison between pseudo-absence selection methods was based on individual model performance and resulting model consensus in a multi-model framework. The full geographic and environmental range of species in the early stages of invasion is usually unknown, especially of those transported globally through trade or tourism. This novel

pseudo-absence selection method will be especially useful for modelling species distributions of invasive species at either a global or regional level.

3.2 Methods

The methods proposed under this chapter are designed to investigate research questions defined by the first two objectives of this thesis. The objectives were: 1. To evaluate the effect of pseudo-absence selection methods both on individual model performance as well as model consensus among different presence-absence models, 2. To develop an improved pseudo-absence generation technique that balances both geographical and environmental space for use in presence-absence correlative species distribution models.

3.2.1 Biotic data

The target species *A. albopictus* and sub-species *D. v. virgifera* were chosen for their different relative occurrence area (ROA) in both geographic (Figure 3.1) and environmental space. Because *A. albopictus* is a critical health hazard, extensive research has been undertaken in areas of insect control, such that there were 3,029 presence points available for this study acquired from literature, personal communication with experts and CABI and GBIF databases (Ikeda *et al.* unpublished data). Out of the 3,029 presence points, 2,928 were spatially unique with respect to the resolution of the environmental data used in this study. For *D. v. virgifera,* there were 64 presence points available for this study (data courtesy of GBIF and PRATIQUE). All *D. v. virgifera* points were used for modelling as they were all spatially unique with respect to the data resolution of the environmental layers.

3.2.2 Environmental data

Data from the BIOCLIM dataset (Hijmans *et al.*, 2005a) which is derived from a 50-year-average (1950-2000) daily temperature and precipitation dataset (WORLDCLIM) (Hijmans *et al.*, 2005b) prepared in 10 arc minute (0.17º) resolution (Hijmans *et al.*, 2005a) was used to source the 19 bioclimatic variables shown in Table 3.1. A geographical variable, elevation, was also obtained through the BIOCLIM data portal. Hijmans *et al.* (2005a) reported that the bulk of the elevation dataset was sourced from NASA's SRTM (NASA-GSFC, 2000) global Digital Elevation Model with additional data from GTOPO30 (EROS, 1996) global elevation data to cover the above 60ºN areas for which there was no SRTM data. Elevation is known to

moderate local climate and it could act as a natural barrier between suitable areas. Elevation was added to account for local topographical variations in habitats.

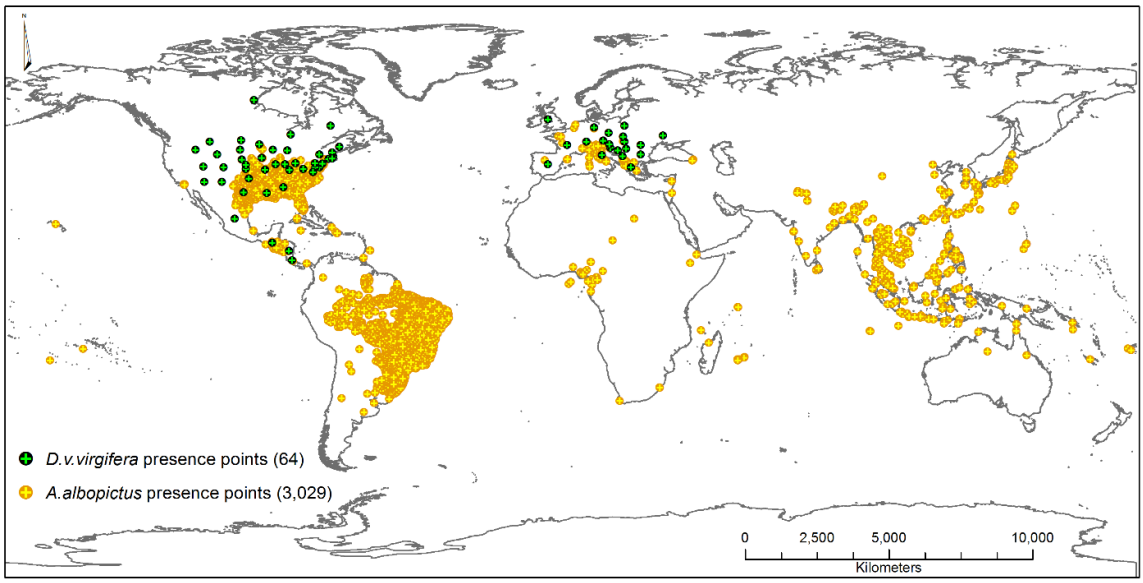


Figure 3.1: Map of global presence data for A. albopictus *and* D. v. virgifera.

Two of the pseudo-absence selection methods in this study use plane circular buffers on background data to limit the pseudo-absence selection within a certain distance from presences. Such planar buffers cannot be overlaid on data in the geographic coordinate system without causing poleward distortion. To avoid this bias, the global (0.17º) and New Zealand extent (30") data were converted into world Mercator WGS 1984 coordinate system and UTM-WGS1984-Zone-59S coordinate system respectively. Both datasets are then resampled into a 15.2 km x 15.2 km and 0.8 km x 0.8 km equal area grids using bilinear interpolation. Optimum cell sizes were determined as follows.

For the global data, the vertical range of the BIOCLIM data (82ºN ~ 56ºS) was used to define latitudinal ranges of 40ºN ~ 40ºS, between 40ºN ~ 60ºN & 40ºS ~56ºS, and greater than 60ºN. The optimum cell size was identified by weighting the average of the mean cell width in each pre-determined latitudinal range by the number of pixels in the latitudinal range. Weighting along latitudinal zones was not necessary for the New Zealand data as the change in horizontal cell size along latitude was small (~0.02 km). The cell size for New Zealand was calculated by taking the square root of the product of the average cell width (0.71 km) and average cell height (0.93 km) in the dataset.

All of the 20 variables were combined in one raster dataset with multiple attributes and converted into a vector point dataset, which was then exported into an ASCII matrix. Each point in the matrix represented an area of 231.3 km^2 within the global data set and an area of 0.64 km^2 within the New Zealand dataset. The total area of analysis covers all global landmass except Antarctica with an area of 135,202,962 km^2. The New Zealand data covered 268,042 km^2.

A non-New Zealand global dataset was used as a background for pseudo-absence selection. This is done to provide all models with a standardized independent dataset (New Zealand) which is used for habitat suitability projections. An environmental similarity test was undertaken by mapping the New Zealand extent in the environmental feature space of PCA transformed BIOCLIM data. There were no New Zealand data points outside the environmental bounds of data for the rest of the world, ensuring models did not extrapolate. The full extent global data was used for global habitat suitability predictions, and the high resolution data was used for habitat suitability projections in New Zealand.

The following sections from 3.2.3 – 3.2.6 give the methods used to obtain four types of pseudo-absence points based on three methods that are currently used and a fourth method proposed in this study. Intermediate results regarding the methods are given within these sections and results that deal with comparison of model predictions based on these methods are given in the result section 3.3.

3.2.3 Simple random pseudo-absence selection (SM1)

Pseudo-absence points are selected randomly from across the whole study area. Known presence points were removed prior to random selection making the size of the background data 134,515,972 km^2 for *A. albopictus* and 135,184,400 km^2 for *D. v. virgifera*. The optimal ratio of presence points to pseudo-absence points that are used to train models is debated (Stockwell & Peters, 1999; Zaniewski *et al.*, 2002; Lobo *et al.*, 2010; Barbet-Massin *et al.*, 2012). An unbalanced design where there are more pseudo-absence points than presence points has been found to affect performance of some models positively, and others negatively (Barbet-Massin *et al.*, 2012). That introduces bias in research designs involving multiple models such as this study. Therefore, an equal number of pseudo-absence points as

presences points were used for the random selection method and all subsequent pseudo-absence selection methods used in this study. Random 2,928 and 64 points were selected for *A. albopictus* and *D .v .virgifera* respectively from the background data.

3.2.4 Spatially constrained pseudo-absence points selection (SM2)

This method uses a spatial constraint on background data before selecting pseudo-absence points. The background data is extracted within a defined distance from presence points. Previous applications of this method have often used an arbitrarily chosen distance VanDerWal *et al.* (2009). Pseudo-absence points were chosen at random from the geographically limited background data. For consistency, in our study the same distances determined within the 3-step method were used. These distances were 350 km for *A. albopictus* and 3,000 km for *D. v. virgifera*. Pseudo-absence points were selected at random from the spatially constrained background dataset. The background data set for this scenario covered 29,219,485 km^2 for *A. albopictus* and 64,791,235 km^2 for *D. v. virgifera.*

3.2.5 Environmental pseudo-absences point selection (SM3)

An environmental profiling, similar to other two-step pseudo-absence generation methods (Chefaoui & Lobo, 2008; Wisz & Guisan, 2009) was performed on the background data except that a one class support vector machine (OCSVM) (Schölkopf *et al.*, 2001) classifier was used. OCSVM is chosen because it can handle high dimensional data and complex non-linear relationships between predictors. The OCSVM model was trained with environmental variable data at presence points. An ensemble of 100 best performing OCSVM models was used to determine robust environmentally profiled background classes (Worner *et al.*, 2014). Using an ensemble approach rather than the single best performing model reduced the probability of choosing an over-fitted model. The OCSVM profiling produced background data with values between zero and one, which represent the probability of being similar to the presence data. All background data points with a probability of 0 (zero-similarity with presences) were extracted as potential pseudo-absence points. Random 2,928 and 64 pseudo-absences were selected from this zero-similarity background data that covered 102,831,933 km^2 and 87,744,064 km^2 for *A. albopictus* and *D. v. virgifera* respectively.

3.2.6 Three step pseudo-absence selection method (SM4)

The novel three-step method developed here provides a balance between using the spatial and environmental space for selection of appropriate pseudo-absence points. The first step is to determine geographic space for the species by establishing the appropriate distance by which background data is bound to presence data. In the second step, an OCSVM model is used to classify the background data constrained in step 1 into various environmental classes. In the third step K-means clustering is used to select a representative sample from all the environmentally dissimilar points identified in step 2 as pseudo-absence points.

3.2.6.1 Step 1: Specifying geographical extent

An independent method based on variable importance analysis was designed to identify an appropriate distance by which background data is bounded to presence points.

Variable importance was analysed using the following steps:

Step 1.1 a baseline data is prepared by extracting the 20 environmental variable values at the presence point locations for each species. Principal component analysis (PCA) was performed on the baseline data. The first *n* principal components (PCs) that make up to 95% of the variance in the baseline data are recorded.

Step 1.2 the loadings (coefficients) of the PCA for each of the selected PCs at Step 1.1 are evaluated and variables that contributed the most (> 70%) to these PCs are noted.

Step 1.3 multiple datasets were produced by bounding background data at different radii from presence points. I chose 50 km, 100 km, 150 km, 200 km, 250 km, 300 km, 350 km, 400 km, and 500 km intervals. PCA was performed on these datasets same as was done for the baseline data. The PCA loadings for the same variables identified at Step 1.2 are extracted over all background datasets. The contribution of these variables versus distance was then plotted and analysed for any decline in contribution to the respective principal components. In cases where no change was observed within the listed intervals the distance was increased by 100 km until change was observed (Figure 3.3). The distance at which the contribution of the most important variables changed was chosen as the optimal limit to bound background data. I suggest that including background data outside the optimum

distance could obscure important information for feature selection. Tuv *et al.* (Tuv *et al.*, 2009) and references therein show that unnecessarily large and redundant background data introduces noise and decreases predictive power of models.

The first three principal components explained 95% of the variance in the background data for both *A. Albopictus* and *D. v. virgifera*. The contribution of the most important variables to their respective principal component declined at 350 km for *A. albopictus* (Figure 3.2), and at 3,000 km for *D. v. virgifera*, these distances were taken as the optimum boundary of background data. The area of the background data extracted from within the optimum distance of presence points was 29,219,485 km² for *A. albopictus* and 64,791,235 km² for *D. v. virgifera*.

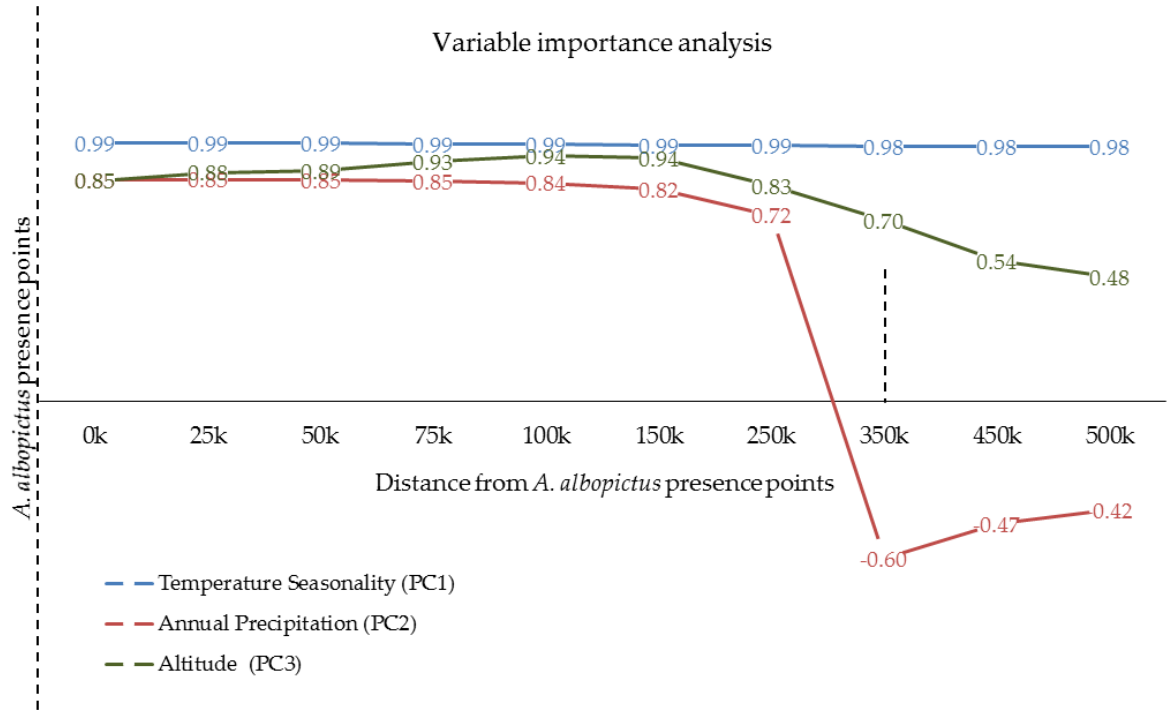


Figure 3.2: Variable importance analysis for A. albopictus *background data delimitation. Graph labels show the coefficients of the three most important variables for* A. albopictus *over the given distances from presence points*

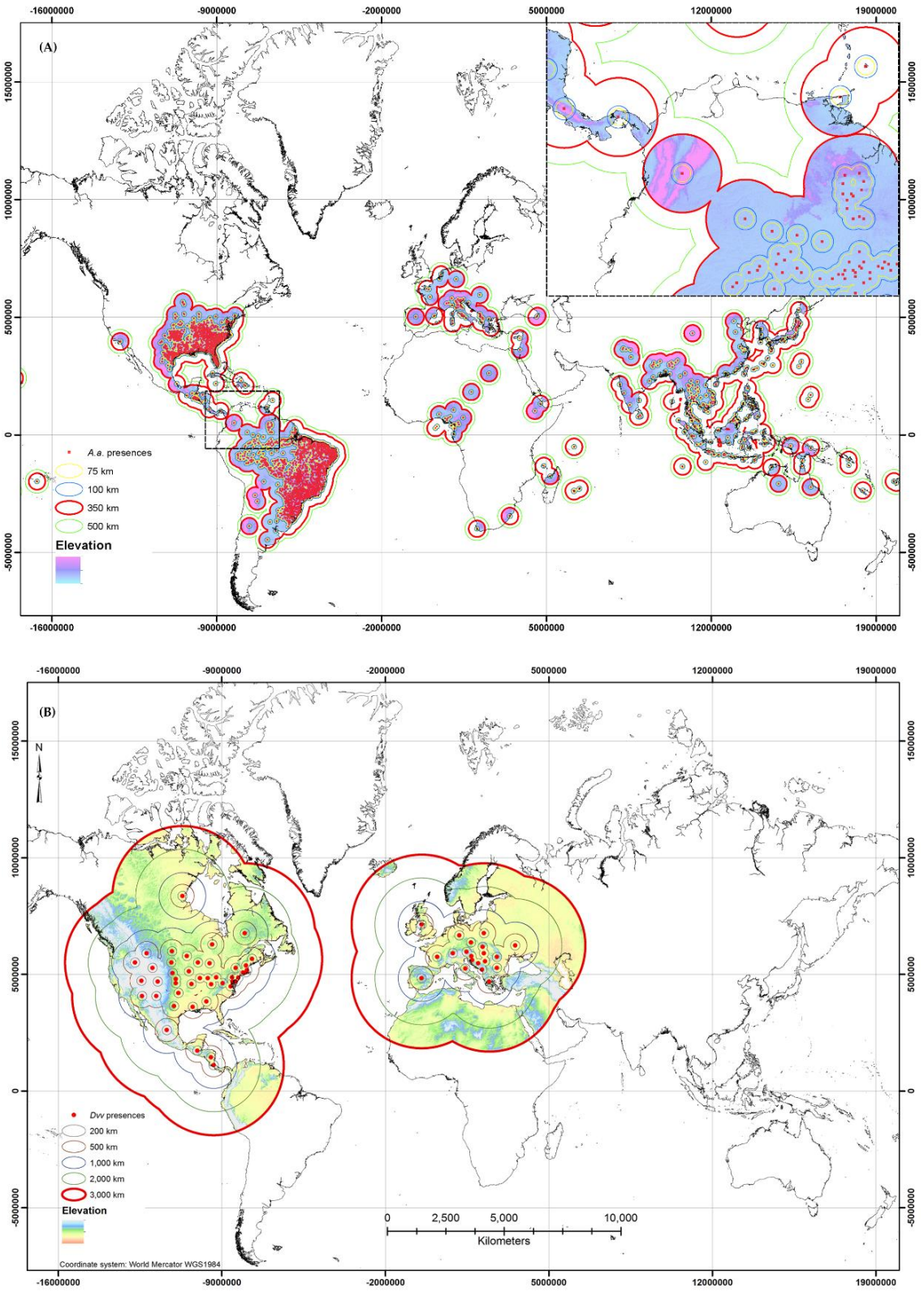


Figure 3.3: Boundaries of background datasets extracted from circular buffers drawn at various radii from (A) A. albopictus & (B) D. v. virgifera presence points.
The bold red boundary shows the optimum background extent identified by the variable importance analysis for the respective case studies.

3.2.6.2 Step 2: Environmental profiling of background data:

Environmental profiling was performed on the spatially limited background data identified at step 1 using an OCSVM (Schölkopf *et al.*, 2001) classifier. All locations with a probability of 0 (zero similarity with presence points) were extracted as a potential background for pseudo-absence selection. This procedure further reduced the background data at step 1 to 9,925,310 km^2 and 12,878,516 km^2 for *A. albopictus* and *D. v. virgifera* respectively.

3.2.6.3 Step 3 K-means clustering

K-means clustering was used to group the zero-similarity locations defined at step 2 into k clusters according to their environmental value as per the specifications recommended by (Worner *et al.*, 2014). The parameter k that determines the number of clusters for K-means clustering was set to the number of presences available (K=2,928 for *A. albopictus* and K= 64 for *D. v. virgifera*). The centroids, from each cluster in the environmental feature space, were selected as they best represented their respective cluster. The projection of the centroids in the geographic space provided the pseudo-absence points needed to proceed with the presence–absence modelling.

3.2.7 Model evaluation and output analysis

The four methods of pseudo-absence selection were compared based on the performance of seven presence-absence models. The seven models were: 1) logistic regression (LOG)(Kleinbaum & Klein, 2005), 2) classification and regression trees (CART)(Venables & Ripley, 1997) 3) conditional trees (CTREE) (Hothorn *et al.*, 2006b). 4) K-nearest neighbours (KNN) (Ripley, 1994); 5) naïve Bayes (NB)(McCallum & Nigam, 1998), 6) support vector machines (SVM)(Vapnik, 1995) and 7) artificial neural networks (NNET) (Haykin, 1998).

Variable selection was carried out using random forests. Random forest (RF) is a classification algorithm that uses an ensemble of classification trees. Random forest is chosen because it is reported to have a good predictive performance even when noisy variables are included (Breiman, 2001). Variable selection was performed independently for each training dataset, as the domain and range of the four types of pseudo-absences vary in the geo-environmental space. Table 3.1 shows the list of variables selected for the different scenarios. For validation, 20% of both the presence and pseudo-absence datasets were partitioned and

set aside for cross-validation while 80% was used to train the models. Each scenario was replicated 20 times. The models were compared based on performance scoring methods (Table 3.2).

The threshold $p > 0.5$ was used to convert model predictions into binary presence-absence maps to obtain predicted presences. There is evidence that shows predefined thresholds such as used in this study may lead to a cut-off that does not approximate the true threshold at which the species is likely to be present (Jiménez-Valverde & Lobo, 2007). The optimum threshold based on prevalence is considered to decrease towards zero for rare species and increases towards one for generalist species (Jiménez-Valverde & Lobo, 2007). Both species used in this research are not rare; thus the bias introduced from erroneous threshold should be similar. The threshold of 0.5 was used as we were interested solely in variation arising from pseudo-absence selection methods. Percentages of predicted presences out of the total study area were compared for differences in habitat suitability predictions among models using the different pseudo-absence methods. Model consensus was analysed for the New Zealand extent, by identifying how many models using the same pseudo-absence method predicted similarly over their respective predicted presence maps.

A habitat suitability prediction was produced using the best model for each pseudo-absence selection scenario at the global extent and for New Zealand. Model Kappa values were used to select the best model for each pseudo-absence method scenario. Kappa is chosen because it corrects for prediction success by chance (Manel *et al.*, 2001). Habitat suitability maps are re-projected back onto a geographic co-ordinate system for visualization. All analyses were carried out using the free software R (R Core Team, 2012) version 2.8.1 and 2.15.1 with packages agricolae (Mendiburu, 2012), class, nnet, MASS (Venables & Ripley, 2002), Coin (Hothorn *et al.*, 2006a), e1071 (Meyer *et al.*, 2007), kernlab (Karatzoglou *et al.*, 2004), klaR (Weihs *et al.*, 2005), multcomp (Hothorn *et al.*, 2008), randomForest (Liaw & Wiener, 2002), SP (Pebesma & Bivand, 2005), VarSelRF (Diaz-Uriarte, 2009). An R based multi-model framework (Ikeda et al, Unpublished data; Worner *et al.*, 2014) was used to run the models in a standardized manner. Data pre-processing and mapping were done using MATLAB version R2011a (MathWorks, 2011) and ArcGIS version 10.1 (ESRI, 2010).

Table 3.1: List of variables selected using 4 pseudo-absence selection methods for the two target species

No.	Variables	aaSM1	aaSM2	aaSM3	aaSM4	dvvSM1	dvvSM2	dvvSM3	dvvSM4
V1	Annual Mean Temperature	✓	✓	✓	✓		✓	✓	
V2	Mean Diurnal Range (Mean of monthly (max temp - min temp))							✓	
V3	Isothermality (P2/P7) (* 100)	✓	✓	✓		✓	✓	✓	
V4	Temperature Seasonality (standard deviation *100)	✓	✓	✓					
V5	Max Temperature of Warmest Month	✓		✓	✓			✓	
V6	Min Temperature of Coldest Month	✓		✓		✓		✓	
V7	Temperature Annual Range (P5-P6)	✓	✓						
V8	Mean Temperature of Wettest Quarter			✓					
V9	Mean Temperature of Driest Quarter	✓		✓					
V10	Mean Temperature of Warmest Quarter	✓	✓	✓	✓			✓	
V11	Mean Temperature of Coldest Quarter	✓	✓	✓		✓		✓	
V12	Annual Precipitation	✓	✓	✓	✓			✓	
V13	Precipitation of Wettest Month	✓	✓	✓	✓			✓	
V14	Precipitation of Driest Month	✓						✓	✓
V15	Precipitation Seasonality (Coefficient of Variation)	✓	✓		✓			✓	
V16	Precipitation of Wettest Quarter	✓	✓	✓				✓	✓
V17	Precipitation of Driest Quarter	✓	✓	✓	✓		✓	✓	
V18	Precipitation of Warmest Quarter						✓		
V19	Precipitation of Coldest Quarter	✓							
V20	Altitude	✓		✓	✓				
Total		17	11	14	8	3	4	13	2

**aa*= Aedes albopictus, *dvv*= Diabrotica v. virgifera, *SM1=random pseudo-absence selection method, SM2 = spatially constrained random pseudo-absence selection method, SM3= 2-step environmental profiling pseudo-absence selection method, SM4 = 3-step environmental profiling with spatial constraint pseudo-absence selection method.*

Table 3.2: Model performance indices

Index	Formula	Abbreviations	Remark
Accuracy	$= \frac{TP + TN}{TP + TN + FP + FN}$	TP=True positive, TN=True negative, FP=False positive, FN=False negative	
Kappa index	$= \frac{(OA - EA)}{(1 - EA)}$	OA = observed agreement(Accuracy) EA = expected agreement EA(TP) $= (TP + FN) * (TP + FP)/(TP + FP + FN + TN)$ EA(TN) $= (FP + TN) * (FN + TN)/(TP + FP + FN + TN)$ $EA = (EA(TP) + EA(TN))/(TP + FP + FN + TN)$	Excellent: >= 0.81 Good: 0.61 – 0.80 Medium: 0.41 -0 .60 Not good: 0.21-0.40 Bad: 0.00-0.20 Very bad: <0.00
Sensitivity	$= \frac{TP}{TP + FN}$		1 - omission error Also referred as "recall"
Specificity	$= \frac{TN}{TN + FP}$		1 – commission error
AUC	ROC curve $TPR = Sensitvity$ plotted against $FPR = \frac{FP}{FP + TN}$	True positive rate vs. False positive rate, Calculated on the test dataset. Threshold cut-off for presence-absence binary prediction is calculated at 50% prevalence for all models.	Values > 0.7 are considered good.

3.3 Results

3.3.1 Pseudo-absences

The environmental range and domain of pseudo-absences from the 4 pseudo-absence selection methods were different for both species (Figure 3.4 for *A. albopictus*). SM1 pseudo-absences had both very close points to presence points as well as environmentally extreme points (Figure 3.4-A). SM2 pseudo-absences were closely clustered around presence points (Figure 3.4-B). SM3 (Figure 3.4-C) and SM4 pseudo-absences points (Figure 3.4-D) were clearly discriminated from presence points. However, the SM4 pseudo-absences did not have environmentally extreme points which is illustrated by their magnitudes on the principal component axes (Figure 3.4).

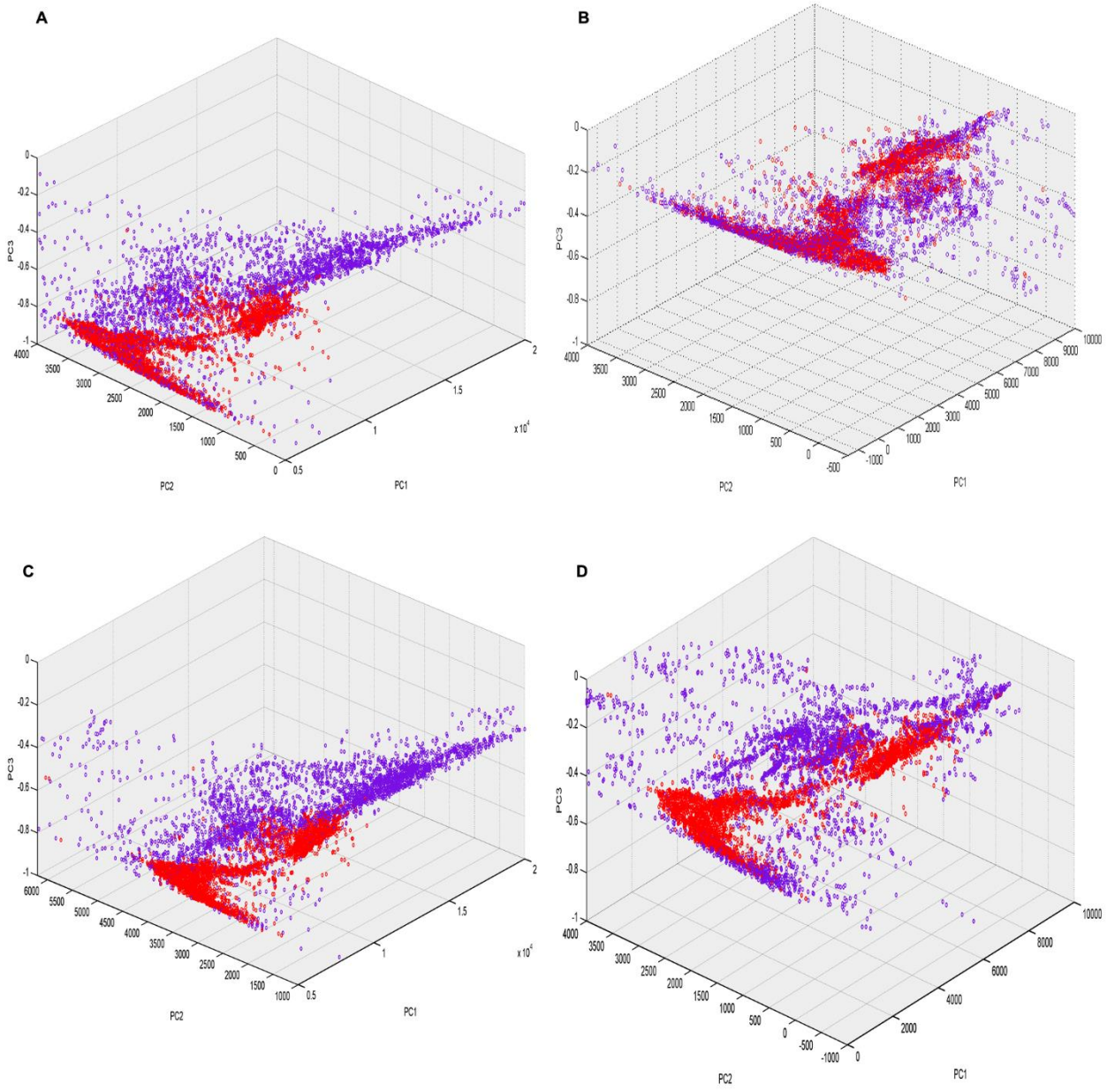


Figure 3.4: Pseudo-absence points from the four pseudo-absence selection methods.
Pseudo-absence points plotted with presence points on the first three principal components of the training dataset for A. albopictus*, (A) SM1, (B) SM2, (C) SM3, and (D) SM4. Red points mark presences and purple points mark absences.*

3.3.2 Variable selection

There was considerable variation in the subset of variables chosen for each training dataset from the total predictor list of 20. The 3-step selection method (SM4) gave fewer variables for both *A. albopictus* and *D. v. virgifera* (Table 3.1).

3.3.3 Model performance

Out of the 56 models from the various data-method-model combinations (7 model types *x* 4 selection methods *x* 2 species), 55 of the models had mean AUC value better than 0.5 meaning all models predicted better than chance except for one model (CTREE,SM1,*Dvv*), which registered a poor performance (AUC = 0.1765). Two-way within subjects analysis of variance was used to calculate the variance attributed to each factor in the experiment. The pseudo-absence selection method had a highly significant (ANOVA, SS[12] = 0.285, p = 0.0017) effect on model mean AUC values, but the interaction between model type and selection method was insignificant (ANOVA, SS = 0.127, p = 0.94). Mean AUC differences due to model type were not significant according to Tukey's HSD test ($p < 0.05$). There was a statistically significant difference in mean AUC of models using SM1 and SM2 pseudo-absences compared with models using SM3 and SM4 pseudo-absences ($p < 0.05$) (Figure 3.5). The average mean AUC of models using SM1, SM2, SM3 and SM4 pseudo-absence points was 0.84 (±0.21 SD), 0.79 (±0.07 SD), 0.95 (±0.05 SD), and 0.95 (±0.03 SD) respectively.

We used the proportion of the sum of correctly predicted pseudo-absences and correctly predicted presences out of the total test data to calculate model accuracy. The ANOVA results for the mean accuracy values for the same models under different pseudo-absence selection methods showed that pseudo-absence selection method has a significant effect on model accuracy (SS= 0.28, $p < 0.0001$). Tukey's HSD test on model accuracy measurements also gave a similar result to comparison of mean AUC values; models using pseudo-absence selection methods SM3 and SM4 have significantly better accuracy than models that used SM1 and SM2 pseudo-absences ($p < 0.05$).

[12] SS= sum of squares

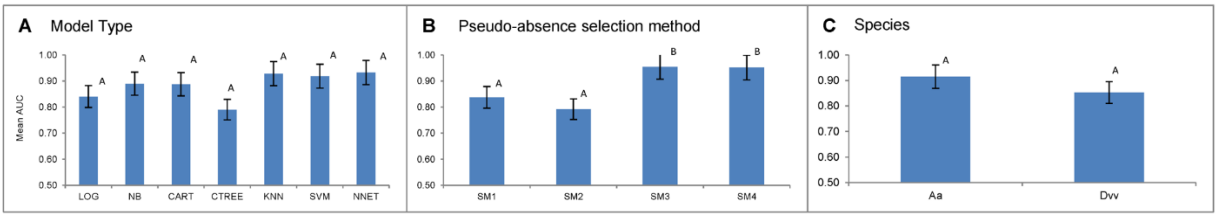


Figure 3.5: Variation on mean AUC values due to model type, pseudo-absence selection method and number and structure of presence data.
Error bars indicate standard errors over replicates. Bars with same letters within a graph are not significantly different (Tukey's HSD test p>0.05). (A) Model type, (B) Pseudo-absence selection method, and, (C) species dataset.

3.3.4 Prediction-reality agreement

The Kappa index was used to compare results between the different models according to pseudo-absence selection method. SM1 resulted in 13 out of the 14 models that were between 'good – bad' bands with the exception of one model (SVM, SM1, *A. a*) in the 'excellent' band (Figure 3.6). The range of scores for the SM1 method was between 0.59 - 0.82 for *A. albopictus* and 0.00 - 0.75 for *D. v. virgifera*. For method SM2, none of the 14 tested models-species combinations were in the 'excellent' band with Kappa values between 0.43 - 0.58 over the two species. For method SM3, model Kappa scores were in the 'excellent' band for 9 out of 14 models and in the 'medium to good' bands for the remaining five models over the two species. For method SM4 model Kappa scores were in the 'excellent' band for 12 out of 14 models and in the 'good' and 'medium' bands for the remaining two models over the two species (Figure 3.6).

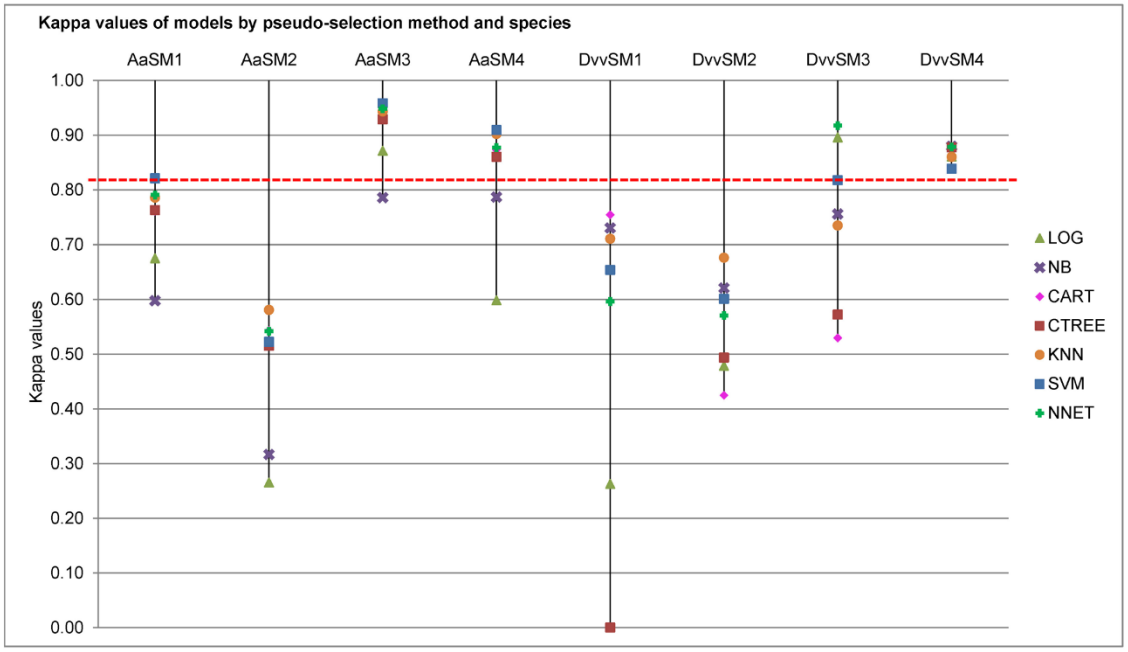


Figure 3.6: Kappa values of models for the four pseudo-absence selection methods and two species dataset.
Aa =A. albopictus, *Dvv* = D. v. virgifera, *values above the red broken line are in the excellent band of the Kappa index.*

3.3.5 Sensitivity and Specificity

Analysis of variance of the specificity results of the seven models showed that there is a highly significant difference between specificity scores of models using different pseudo-absence selection methods (SS= 0.54, $p < 0.0001$) over the two species, and the model type also had a significant contribution towards the variation in the specificity results (SS = 0.14, $p = 0.011$). The lowest mean specificity values were obtained from models using pseudo-absence selection method SM2, models using SM1 pseudo-absence points also had low specificity scores but were significantly better than SM2 models (Figure 3.7). Models that used SM3 and SM4 pseudo-absence points gave significantly better specificity than SM1 and SM2. There was a similar trend for sensitivity where the pseudo-absence selection method had a significant effect on model sensitivity (SS = 0.095, $p = 0.025$). All models with SM3 and SM4 pseudo-absences scored high sensitivity values (> 0.85) for both species dataset (SM3, mean = 0.90, SD = ±0.10; SM4, mean = 0.91, SD = ±0.02). While models with SM1 and SM2 pseudo-absences had low sensitivity scores (SM1, mean 0.85, SD = ±0.14; SM2, mean 0.81, SD

= ±0.05). There was a considerable between-species variation with respect to sensitivity scores of models using SM1 and SM2.

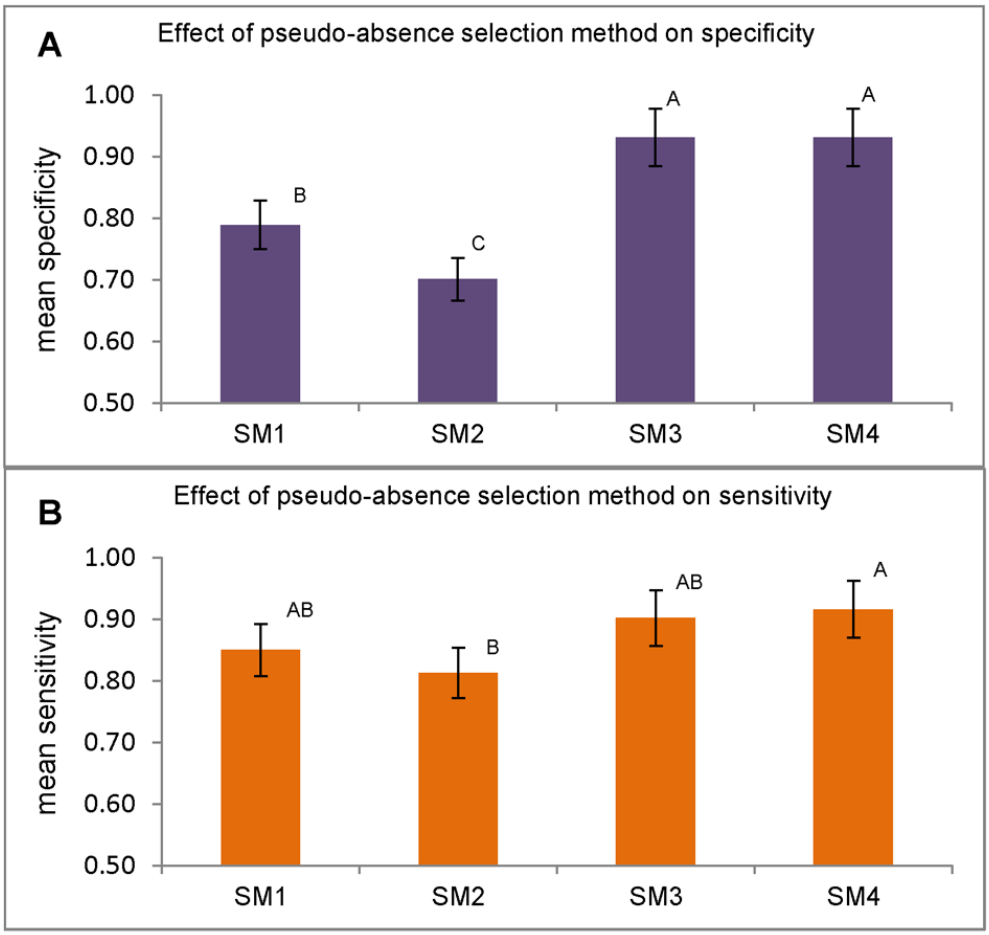


Figure 3.7: The effect of pseudo-absence selection method on mean specificity and sensitivity values. Error bars indicate standard errors. Bars with same letters are not significantly different (Tukey's HSD test p>0.05), (A) specificity (B) sensitivity.

3.3.6 Predicted prevalence, model consensus and habitat suitability

There were variations with respect to the number and location of predicted presences by models that used the four different pseudo-absence selection methods. For the global analysis, models using the SM4 method resulted in the highest percentage of predicted presences (mean = 39.29%, SD = ± 17.65), with SM2 and SM3 ranking second (mean = 31.70%, SD = ±18.24) and third (mean = 24.82%, SD = ±10.19) respectively while SM1 (mean = 22.77%, SD = ±11.15) gave the smallest percentage of predicted presences. For the New Zealand data,

methods SM2 (mean = 52.42%, SD = ±30.27), SM3 (mean = 50.30%, SD = ±37.11), and SM4 (mean = 51.81%, SD = ±29.68), gave very similar predicted presence percentages. The percentage of predicted presences from models using SM1 pseudo-absences was significantly lower (mean = 9.90 %, SD = ±17.78) than models using all the other three methods (p= 0.01, p = 0.02, p = 0.01, Tukey's HSD test in comparison with SM2, SM3 and SM4 respectively).

The predicted presences for both *A. albopictus* and *D. v. virgifera* in New Zealand were analysed to investigate the level of model consensus in the predictions. Model consensus was categorized as follows; prediction by one model = no consensus, prediction by two models = low consensus, prediction by 3-4 models = moderate consensus, and prediction by 5-7 models = high consensus. Predicted presence percentages and model consensus levels are given in figure 3.8.

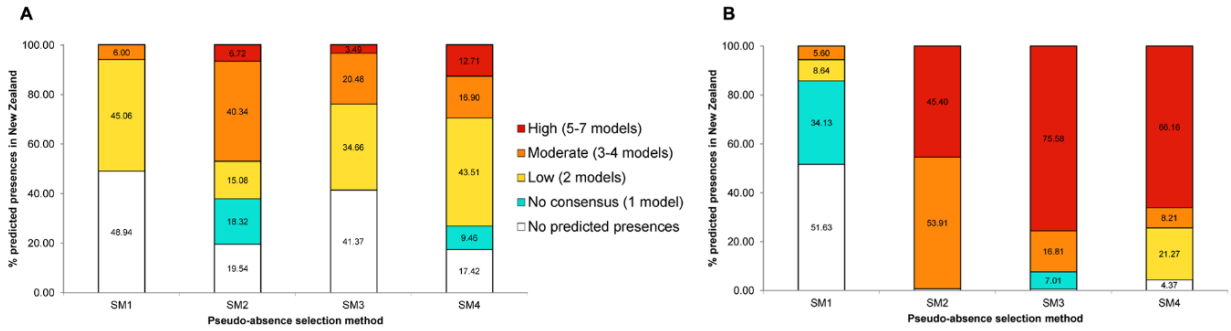


*Figure 3.8: Percentages of predicted suitable areas and respective model consensus on predictions in New Zealand. (A), Asian tiger mosquito (*A.albopictus*) (B), Western corn rootworm (*D. v. virgifera*)*

Habitat suitability maps were produced using the best models, according to Kappa score, for each scenario. For the *A. albopictus* dataset, the best performing models based on SM1, SM2, SM3 and SM4 pseudo-absence methods were NNET, KNN, NNET and SVM respectively. For the *D. v. virgifera* dataset, the best performing models based on SM1, SM2, SM3 and SM4 pseudo-absences methods were NNET, NB, CART, and KNN respectively (Figures 3.9 & 3.10).

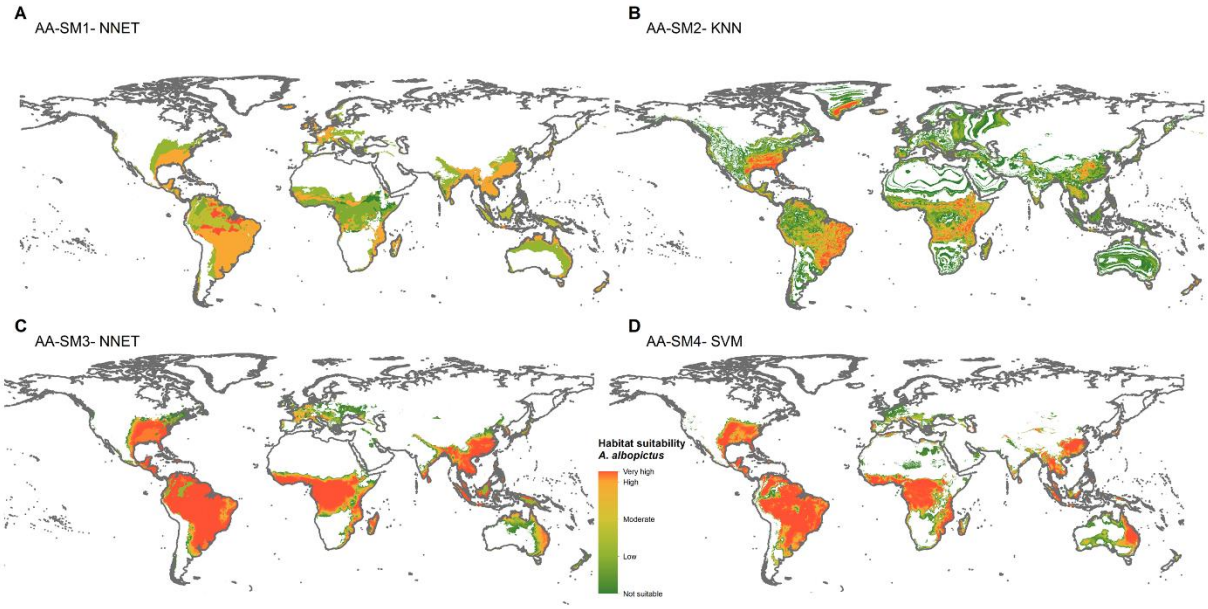


Figure 3.9: Global habitat suitability prediction for Asian tiger mosquito (A. albopictus). (A), SM1 pseudo-absences with model NNET (B) SM2 pseudo-absences with model KNN (C), SM3 pseudo-absences with model NNET (D) SM4 pseudo-absences with model SVM. Note: A.albopictus occurrence data is too dense to overlay on prediction, refer to Figure 3.1.

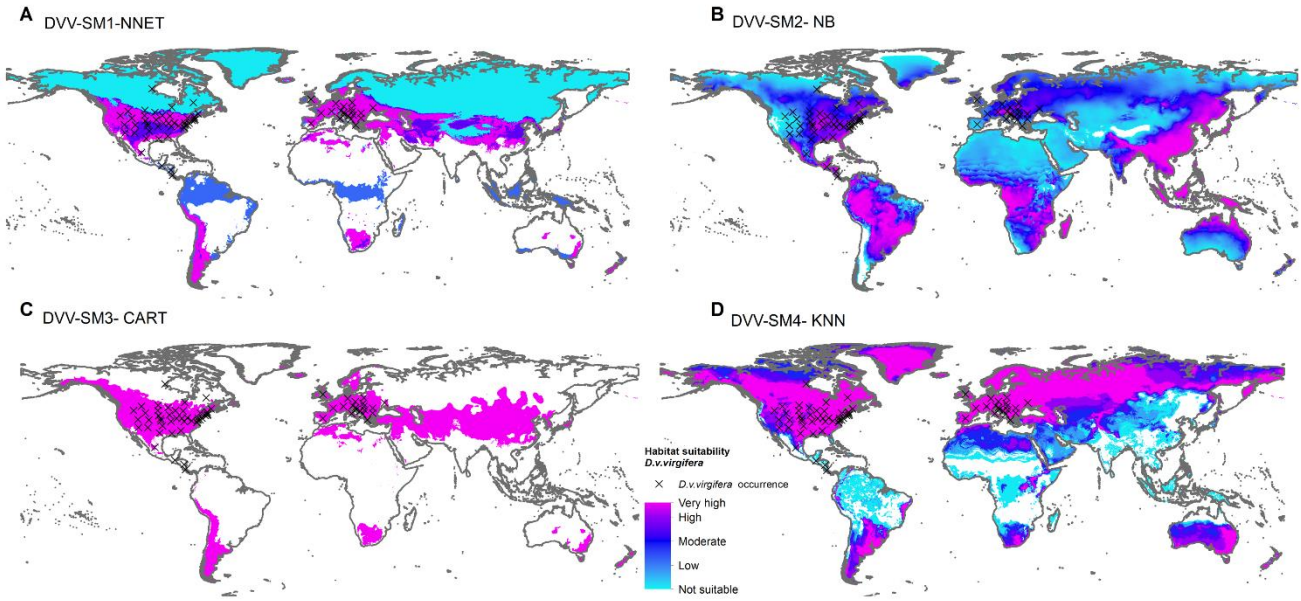


Figure 3.10: Global habitat suitability prediction for Western corn rootworm (D. v. virgifera) *SM1 pseudo-absences with model NNET (B) SM2 pseudo-absences with model NB (C), SM3 pseudo-absences with model CART (D) SM4 pseudo-absences with model KNN.*

Habitat suitability map comparisons in the projected range (New Zealand) show that SM1 based maps were dissimilar from SM2, SM3 and SM4 suitability maps (Figures 3.11, 3.12). The SM1 suitability predictions both for *A. albopictus* and *D. v. virgifera* in New Zealand were

limited to very small areas of low to moderate suitability. The habitat suitability projected using SM2 pseudo-absences identified 72,557 km^2 of highly suitable area (> 0.9 probability) for *A. albopictus* and 92,779 km^2 of highly suitable area for *D. v. virgifera*. The suitability prediction based on SM3 pseudo-absences identified no highly suitable locations for *A. albopictus* and a large 247,883 km^2 area of highly suitable area for *D. v. virgifera*. Habitat suitability prediction based on the 3-step method (SM4) identified 8,752 km^2 of highly suitable area for *A. albopictus* and 151,569 km^2 for *D. v. virgifera.*

3.4 Discussion

A number of studies have established that the pseudo-absence selection method used for SDMs affects model performance (Thuiller *et al.*, 2004; Chefaoui & Lobo, 2008; Jiménez-Valverde *et al.*, 2008; Lobo *et al.*, 2010; Barbet-Massin *et al.*, 2012). In this study, the effect of pseudo-absence selection methods on the performance of seven models was investigated. The results showed that methodological prescription of pseudo-absence points, similar to the 3-step method developed in this study, enhances model predictive power. The commonly used approaches are to constrain the background data geographically (similar to SM2), or environmental profiling of the background data (similar to SM3) (Zaniewski *et al.*, 2002; Chefaoui & Lobo, 2008; Barbet-Massin *et al.*, 2012). However, some studies have reported that random pseudo-absence selection method (equivalent to SM1) works best in some contexts. For example, SM1 is considered to work well with logistic regression models (Barbet-Massin *et al.*, 2012) and when environmental data is too complex to perform environmental profiling (Wisz & Guisan, 2009). Jiménez-Valverde *et al.* (2008) and Lobo *et al.* (2010) suggested that the best way to get potential distribution representation of a species is by using absences located relatively near the external boundary of the environmental domain and adding geographic proximity if the requirement is to get the realized distribution representation. In the three-step method, we quantified these boundaries by utilizing variable importance analysis over various distances from presence locations. The challenge was to maintain model performance while introducing spatial constraint on the potential background data. Environmentally profiled background data without any

geographical constraint usually gives very high model AUC and sensitivity values because the data are overly and unrealistically discriminated. Rather than using an arbitrary distance, the 3-step pseudo-absence selection method utilizes an ecologically meaningful distance to specify geographic extent of background data, in order to minimize information loss due to the introduced spatial constraint. I found the optimum distance for the background data extent to be 350 km for the *A. albopictus* dataset and 3,000 km for *D. v. virgifera* dataset. Care should be taken not to associate distance obtained through variable importance analysis as a constant biogeographic characteristic of the species. The distance at which background data is bounded is identified based on the species relative area of occurrence. As a consequence, it is affected by the number of presence locations, their distribution and the extent of the study area. The identified distance must be re-calculated if the presence data or the extent of the study area changes.

3.4.1 Variable selection

Variable selection is an essential step in species distribution modelling. Selected variables and their relationship at the presence points are the mechanism by which ecological assumptions are incorporated in correlative species distribution models. Failing to select the appropriate explanatory variables leads to model results detached from ecological reality. In this study, we found large variation between the numbers and types of variables selected according to presence data and pseudo-absence selection method.

The between-species differences in the variables selected for each pseudo-absence scenario can be used to assess the effect of species presence data on variable selection. More variables were selected for the *A. albopictus* training dataset than *D. v. virgifera* in all pseudo-absence selection methods. This was because the *A. albopictus* dataset with 2,928 presence points covers a large area in geographic and environmental space, requiring more variables to characterise the training data than the *D. v. virgifera* dataset that has 64 presence points over a relatively limited geographic and environmental range. This result is not unexpected, the larger the environmental range of the species, the larger number of variables needed to construct a valid model.

The within-species differences in the variables selected show that pseudo-absence data has considerable influence on variable selection. A large number of variables in this case correspond to inconsistent pseudo-absence points that require a large number of variables to characterise the training data. The least number of variables were selected from data using the 3-step method (Table 3.1). More conservative variable selection is a result of a unique interplay of limiting background extent and robust environmental profiling used in the 3-step method, which excluded environmentally extreme outliers in the training data while providing clear environmental classification between presence and pseudo-absence points.

It is well established that the number of presences and the environmental data are critical for variable selection and accuracy of SDM predictions. However, defining appropriate unsuitable areas by selecting optimal pseudo-absences to contrast with suitable areas inferred from presence points is equally important. Moreover, if certain methods have comparable performance, it is better to select the one with less number of variables for a simpler model that can easily be interpreted (Beaumont *et al.*, 2005; Aragón *et al.*, 2010).

3.4.2 Model performance

With respect to model Kappa values, SM1 results show that random pseudo-absence selection method is not consistent either for the two species or the seven models tested. For example, the logistic regression model (LOG) performed well for *A. albopictus* with a high Kappa value but performed poorly for *D. v. virgifera*. This inconsistency is confirmed by Lobo *et al*. (2010) who states that random pseudo-absence selection methods are unreliable due to their high dependence on species presence point distribution and abundance. High model performance using this method can occur by chance and is unlikely to be repeatable for different species or model scenarios as shown in this study. SM2 results were low for all models. Both SM1 and SM2 resulted in significantly low mean AUC and specificity scores compared with models using SM3 and SM4 pseudo-absences. SM1 and SM2, therefore, seem not ideal pseudo-absence selection methods to use in SDMs.

SM3 gave consistently high model performance (Kappa statistics) except for CTREE and CART models which had variable performance across the two species. The machine learning models using SM3 pseudo-absences performed consistently over the two species dataset.

SM3 was found to perform well, especially for the LOG model giving similar high Kappa values for both species. This result is despite reports stating that regression models work best under random selection methods (Wisz & Guisan, 2009; Barbet-Massin *et al.*, 2012). I attribute the good results from the LOG model on SM3 pseudo-absences to the use of a robust model (OCSVM) for environmental profiling of background data.

SM4 provided excellent Kappa values for all models for the *D. v. virgifera* data set and five models of *A. albopictus* dataset. A single low Kappa value was reported for the LOG model performance. There was no significant difference between AUC, sensitivity and specificity values between SM3 and SM4 methods despite that the background data for the pseudo-absence points of SM4 were geographically restricted. While there was no statistical difference, SM4 method achieves high model performance while avoiding extreme spatial and environmental locations that could lead to inconsistency in prediction for new areas.

3.4.3 Model consensus and habitat suitability

The highest percentage of predicted presences was obtained from the 3-step pseudo-absence selection method. This result is very important especially for invasive species studies where identifying potential areas suitable for the establishment for the target species is critical. The lowest predicted presence percentage was from the random selection method (SM1) both at a global and New Zealand scale. Comparisons of predicted presence maps were done to check consensus among models that used the same pseudo-absence method. I recognize that model consensus alone does not ensure high prediction accuracy because models can wrongly agree on the occurrence of a species. A good example is the high consensus among models using SM2 pseudo-absence points for prediction of *D. v. virgifera* distribution in New Zealand (Figure 3.8-B), even when the Kappa model performance scores for these models were very low. However, high model consensus combined with high model performance scores is preferable to multiple models with high performance scores and low agreement. Furthermore, inconsistency between predictions makes SDM result interpretations difficult for decision makers. In this study, the three step method (SM4) provided the needed combination of high model performance in terms of Kappa values (Figure 3.6) and consistency in model predictions in terms of high model consensus (Figure 3.8-A, B).

Habitat suitability predictions based on the 4 pseudo-absence types gave different results in terms of the size and location of suitable areas for *A. albopictus* and *D. v. virgifera*. Pseudo-absence points from SM1 and SM2 methods are not distinctly separated from presences in the environmental feature space (Figure 3.4-A, B). This lack of discrimination is reflected in their respective habitat suitability predictions. Both SM1 and SM2 maps showed underestimation of the potential suitable area *for A. albopictus* and *D. v. virgifera* when overlaid with occurrence points. Pseudo-absences from both SM3 and SM4 methods were distinctly clustered away from presence points in the feature space allowing environmental discrimination (Figure 3.4-C, D). Accordingly, most of the occurrence areas are identified by the SM3 and SM4 models as highly suitable for both species. While such high model sensitivity is beneficial to more accurately estimate the potential distribution of a species, it is possible to overestimate the potential distribution if highly discriminated presence/pseudo-absence training data are used (Lobo *et al.*, 2010). Therefore, even if both SM3 and SM4 gave comparable suitability predictions, it is advisable to determine optimum background extent for pseudo-absence selection if the study area is at a global or regional scale.

3.4.4 Implications for future **A. albopictus** *and* **D. v. virgifera** *management in New Zealand*

Aedes albopictus: the global distribution estimated for *A. albopictus* from SM1 and SM2 appropriately covered the native Southeast Asian and the introduced South American range, but did not cover the North American distribution accurately. The European and African population were also not accurately represented on the maps (Figure 3.9 –A, B). SM3 and SM4 global distribution maps for *A. albopictus* reflect the current complete range of *A. albopictus*. However, the extent of predicted suitable areas for *A. albopictus* in New Zealand varies between projections using SM3 and SM4 pseudo-absence methods. The SM3 projection (Figure 3.11-C) only shows 2,000 km^2 of moderately suitable area within New Zealand, whereas the SM4 projection identified over 8,000 km^2 of highly suitable areas (Figure 3.11-D).

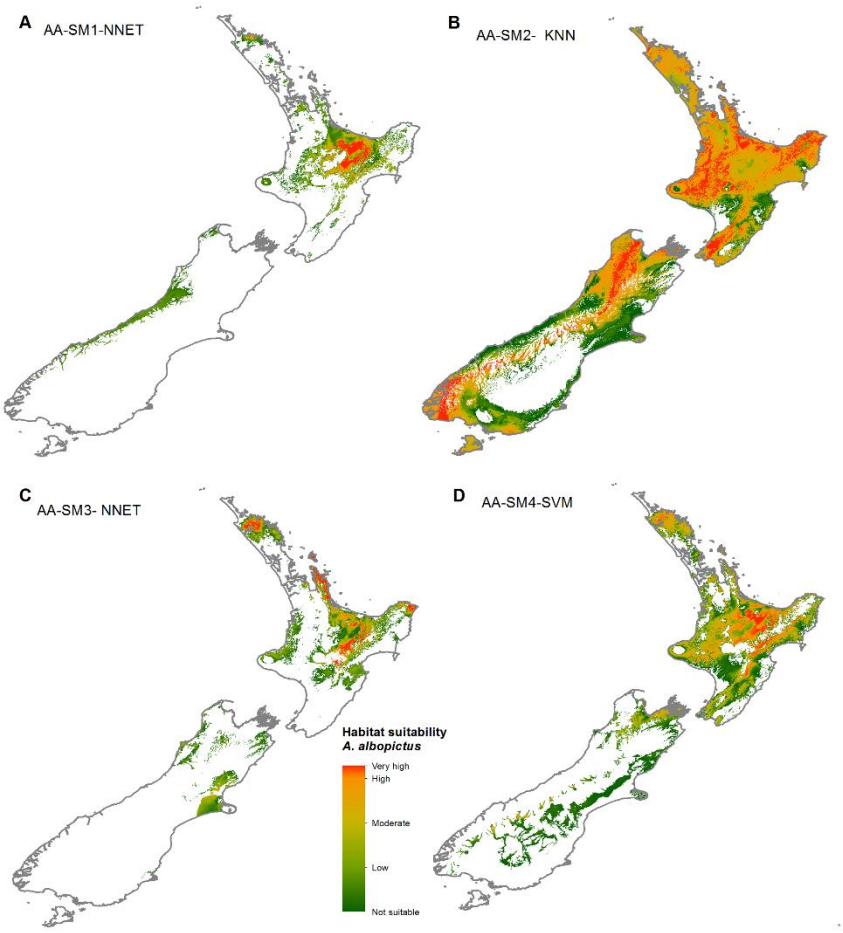


Figure 3.11: Habitat suitability prediction for Asian tiger mosquito (A. albopictus) in New Zealand. (A), SM1 pseudo-absences with model NNET (B) SM2 pseudo-absences with model KNN (C), SM3 pseudo-absences with model NNET (D) SM4 pseudo-absences with model SVM

Given that other species from the *Aedes* genus have established in New Zealand and that *A. albopictus* is repeatedly intercepted at the New Zealand border (Derraik, 2004), I suggest that the suitable areas identified by SM4 be considered in future mosquito related biosecurity assessments. The suitability projection difference between the SM3 and SM4 shows that incorporating a spatial dimension while environmental profiling has a significant effect on model predictions. *A. albopictus* is a particularly difficult species to model as it is currently undergoing a rapid range expansion. Previous studies showed that there is a niche shift throughout the dispersal history of *A. albopictus* (Medley, 2009). It is important to select accurate presence and pseudo-absences data while projecting suitable areas for such species whose distribution spans a wide environmental range.

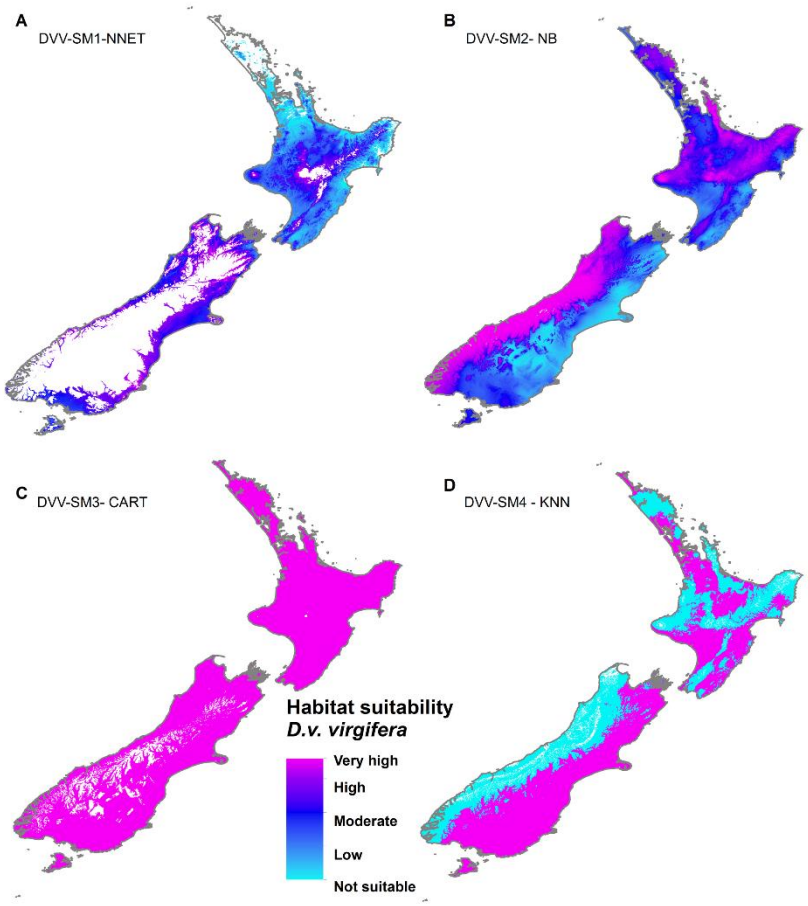


*Figure 3.12: Habitat suitability prediction for Western corn rootworm (*D. v. virgifera*) in New Zealand.*
(A), SM1 pseudo-absences with model NNET (B) SM2 pseudo-absences with model NB (C), SM3 pseudo-absences with model CART (D) SM4 pseudo-absences with model KNN

Diabrotica v. virgifera: similar to *A. albopictus*, the SM1 and SM2 global species distribution model for *D. v. virgifera* did not fully reflect the current known distribution of the species (Figure 3.10-A, B). The SM3 and SM4 predictions (Figure 3.10-C, D) reflected the current known distribution, although the former was more conservative and the latter failed to characterize Central America, the native habitat of the species as highly suitable. I presume this under-prediction of the Central American range is not related to the model and pseudo-absence selection method used because similar result was obtained in previous studies that utilized a variety of modelling approaches (Aragón *et al.*, 2010; Dupin *et al.*, 2011; Kriticos *et al.*, 2012a). An interesting variation in prediction of SM4 is the highly suitable areas identified close to East Africa, an area into which *D. v. virgifera* is expected to spread unless appropriate prevention measures are taken (Hummel *et al.*, 2008). The SM4 suitability

projection for *D. v. virgifera* in New Zealand showed northern and central areas of the North Island and areas east of the Southern Alps as highly suitable (Figure 3.12- D). Although maize (*Zea mays*) production is not a major economic crop in New Zealand, it still accounts for 30% of the arable industry (Barber *et al.*, 2011). Biosecurity measures at the border are essential to prevent the entry of *D. v. virgifera*, a major maize pest, to New Zealand.

3.4.5 Does model type matter?

Several studies show that model type is a major source of uncertainty in SDM results (Dormann *et al.*, 2008; Buisson *et al.*, 2010) among other factors like variable selection, data collinearity and pseudo-absence selection. Uncertainty in SDMs can also arise both from data inaccuracy and internal model error (Elith & Graham, 2009). While little can be done by users to fix errors inherent in model algorithms, model error from data inaccuracy can be reduced by boosting input data quality. Models perform differently given different datasets (environmental data, presence data and pseudo-absence data). While the effect of the accuracy of environmental and presence data have been investigated in depth, the effect of accuracy of pseudo-absence points on model performance has been less investigated. In this study, I established that a robust pseudo-absence selection method can create an input dataset that improves the performance of the SDMs investigated here. That is shown by the low standard deviation in model results that used the 3-step (SM4) pseudo-absence points and the very high Kappa values. Well-structured training data with appropriate variables increases the performance of all models. However, it is still very important to choose models carefully while keeping presence data quality, environmental data and model expertise in mind.

3.4.6 Caveats

The first step of the 3-step pseudo-absence selection system that identifies the appropriate distance within which background data is to be extracted can be quite time consuming and tedious. This can be overcome by developing an automated framework to test variable importance at a set of pre-set intervals.

Another concern is that when large number of presence points coincide with a small background extent in step 1. A small background extent that encompasses a large number

of presence points may reduce the area available for environmental profiling at step 2. That could lead to a poorly discriminated environmental classification. That is not expected to be a common problem as accurate presence points are not usually available in abundance at a global or regional level. This, however, could be remedied by introducing a threshold that relates density of presence points to a minimum distance at which spatial extent of the background data is drawn.

3.5 Summary

When the complete range of a species is unknown, visualizing the distribution of the known presence locations both in geographic and environmental space and assessing the species ROA, is valuable. If presence data is highly clustered both in geographical and environmental space, using presence-only models often leads to extrapolation. In such cases, it is advisable to use presence-absence models with a pseudo-absence selection method that considers both the spatial and environmental space (Jiménez-Valverde *et al.*, 2008; Lobo *et al.*, 2010). When performing species distribution modelling for species undergoing rapid range expansion with dynamic presence data records, new distances should be re-calculated to specify background data geographic extent with the addition of new presence points according to variable importance analysis over various distances from the new presence dataset.

The three-step pseudo-absence selection method (SM4) was shown to result in high model performance while spatially constraining background data to filter out extreme geographically dissimilar locations. Any loss of information from bounding background data geographically before environmental profiling is compensated by the added precision resulting from reduced over-fitting of an SDM model. While this result holds for the models tested in this study, further investigation over more species and models is recommended.

Chapter 4

4. Why do models predict differently for the same species and/or locations?

4.1 Introduction

Various habitat suitability models have been used to predict the spatial distributions of a number of invasive species. Many such studies are based on correlative modelling that use species presence points along with environmental data to infer suitable habitats for species under study (Elith & Leathwick, 2009). Currently, discrepancies between model results have become a major issue in the field of ecological modelling. Overall concerns regarding discrepancies among model results and especially the need for quantification of uncertainty of model results have been repeatedly discussed in the literature (Elith *et al.*, 2002; Thuiller, 2004; Araújo & Guisan, 2006; Elith *et al.*, 2006; Hartley *et al.*, 2006; Pearson *et al.*, 2006; Araújo & New, 2007; Dormann *et al.*, 2008; Buisson *et al.*, 2010; Venette *et al.*, 2010).

The difference in prediction accuracy of different models is usually attributed to the inherent robustness of model algorithms. Simple models (for example BIOCLIM[13], DOMAIN[14]) are reported to be good for prediction of rare species with a limited environmental range representing uncomplicated interaction among environmental variables, while complex

[13] BIOCLIM: a bioclimatic species distribution model using rectilinear polygons to profile environmental data based on presence points (Busby, 1986)

[14] DOMAIN: a bioclimatic species distribution model using disconnected convex hulls to profile climatic data based on presence points (Carpenter *et al.*, 1993)

models (example SVM[15], ANNs[16]) with complex functions that consider non-linearity and large number of variables can handle complex interactions within a high dimensional variable space (Elith *et al.*, 2006; Tsoar *et al.*, 2007). Jiménez-Valverde *et al.* (2008) and Lobo (2008) argued that the above claim is not entirely true due to inappropriate comparison of model prediction results for species with varying relative occurrence areas. However we can still conclude that there is an inherent difference in the predictions between simple and complex models from the difference between model predictions in studies that compared models based on the same occurrence data and study area (Senay *et al.*, 2013).

(Figure removed, subject to copyright)

Jiménez-Valverde, A., Lobo, J. M., & Hortal, J. (2008, p2). Not as good as they seem: the importance of concepts in species distribution modelling. Diversity and Distributions, 14(6), 885-890. doi:10.1111/j.1472-4642.2008.00496.x

Figure 4.1: A hypothetical gradient of potential – realized geographic species distribution outputs aligned with species distribution modelling methods of varying complexity.

The type and number of variables used further complicate the gradient between simple and complex models (Figure 4.1) because of the effect of variable interactions which differ among different variables. How well the interactions among environmental variables is represented in models affects the prediction accuracy of both simple and complex model

[15] SVM: (Support vector machines) a machine learning algorithm based on artificial neural networks, can handle large number of variables and non-linearity, uses a hyper-plane classifier (Vapnik, 1995)

[16] ANNs: (Artificial neural networks) a machine learning network that mimics human neurons, the different nodes in the model adjust learning by back-propagating errors from earlier time steps in the learning process (Haykin, 1998)

types (Jiménez-Valverde *et al.*, 2008; Buisson *et al.*, 2010). Complexity of variable interactions makes choice of predictor data one of the fundamental components of species distribution modelling. The use of geo-environmental variables in addition to purely climatic variables like temperature and precipitation have been reported to increase prediction accuracy in SDM models (Pereira & Itami, 1991; Zimmermann *et al.*, 2007; Elith & Leathwick, 2009; Kearney & Porter, 2009).

However many studies still depend on using limited variables, for example variables derived only from temperature and precipitation data without additional climatic or geo-environmental information for example, elevation that might help models discriminate an ecological niche better (Austin & Van Niel, 2011). Often the spatial variation of geo-environmental variables is much higher than the climatic variables, and can help characterize unique habitats when used along with climatic variables (Zimmermann *et al.*, 2007).

General reluctance to use additional environmental variables may be associated with the data inconsistency that might occur when using multi-sourced and multi-scale variables. Because, as the types and number of variables increase the likelihood of getting variables from the same source decreases. Moreover, scale of the multi-source variables will also likely differ, which increases modelling effort and becomes relatively harder to handle. The other major complication with increased number of variables is increased complex interactions among large set of variables. Climatic variables that have been frequently used in a number of species distribution studies are assumed to have linear relationships which made using simple models possible. However, ecological theory does not generally support that one can always assume linear relationships between environmental variables (Austin, 2007), and this is even more apparent if a large number of predictors from multiple data sources and multiple scales are used in the modelling process. Therefore choice of predictor data is a factor that further affects model prediction accuracy in addition to internal model robustness.

Finally, any dimension reduction performed on a predictor dataset may also affect model prediction accuracy. Numerous dimension reduction methods have been used for various

applications but only a few have been successfully adopted in ecology (Luoto *et al.*, 2004; Heikkinen *et al.*, 2005; Dormann *et al.*, 2008). Nonetheless, as more GIS and remote sensing data are becoming available for species distribution models improved or more appropriate dimension reduction methods also need to be adopted. When using multi-sourced environmental predictor variables in combination with climatic variables, it is important to consider the possibility of complex non-linear interactions between variables before choosing a dimension reduction method. For instance, a high dimensional predictor dataset that may not be adequately handled by a simple model can be modelled after application of appropriate dimension reduction methods (Guyon & Elisseeff, 2003).

Few studies have investigated sources of uncertainties in SDMs (Hartley *et al.*, 2006; Pearson *et al.*, 2006; Dormann *et al.*, 2008; Buisson *et al.*, 2010) and even fewer have developed techniques to quantify the uncertainties associated with modelling in ecology and the wider spatial modelling context (Wang *et al.*, 2005; Hartley *et al.*, 2006; Yemshanov *et al.*, 2012). Based on the recommendations from previous studies, which are mainly to develop frameworks that report model prediction uncertainty, this study investigates whether it is possible to prescribe a predictor list, dimension reduction methods, and models based on the spatial distribution of presence points, and their pattern in the predictor feature space. If it is possible to recommend a set of conditions prior to developing species distribution models by examining the pattern of presence and associated absence points in the predictor feature space, uncertainties in model predictions derived from using inappropriate tools can be avoided.

4.2 Methods

Four of the objectives in this thesis are covered by the research conducted in this chapter. The objectives were: 1) to determine if a wider range of geo-environmental predictors additional to the commonly used temperature and precipitation predictors improve global and regional insect habitat suitability predictions, 2) to investigate the effect of linear and non-linear dimension reduction methods on species distribution model performance, 3) to investigate the effect of model component interactions on species distribution model performance based on a factorial experiment and selected case studies, 4) to develop species

distribution prediction assessment indices that complement confusion matrix based validation methods. The research design and methods used to conduct studies in line with the above objectives are given below.

4.2.1. Research design and model conceptualization

The study was carried out using a 3 x 3 x 4 x 5 factorial design to test the effects of predictor choice and data processing and analysis in habitat suitability modelling. The design incorporated three types of predictor data, three kinds of collinearity reduction methods, four types of models that utilize different techniques of modelling; and five species. Species results were considered as replicates to understand variation in accuracy prediction sourced from occurrence data.

The presence datasets from the five species (SP1, SP2, SP3, SP4, SP5) were first used to generate nine sets of pseudo-absence datasets (for each species) based on the three predictor datasets (P1, P2, P3) and three dimension reduction method (DR1,DR2, DR3) combinations available (Figure 4.2). After pseudo-absence data selection, nine training/test datasets that have equal number of presence and pseudo-absence data were prepared for each species, totalling 45 training/test datasets across the five species. The ratio of test-train data was set at 1 to 5 where 20% of the data was used for testing models and 80% for training models, 20 replications were carried out to take the average model performance for each train-test cycle. The training/test datasets of each species were used to train and evaluate four model types (MT1, MT2, MT3, MT4) and finally predict global species distributions of the five species according to the different predictor-dimension reduction-model type combinations.

With this layout each species had 36 different species distribution predictions which is a product of combinations of three predictor datasets, three dimension reduction methods and 4 model types. Five different confusion-matrix-based indices and one model generated (Table 4.3) performance index were used to assess the cross-validation results. The best and worst predictions were compared to highlight the best and worst data-method-model combinations for each species. The repercussions of predictor data choice and associated data pre-processing techniques on model prediction accuracies were discussed.

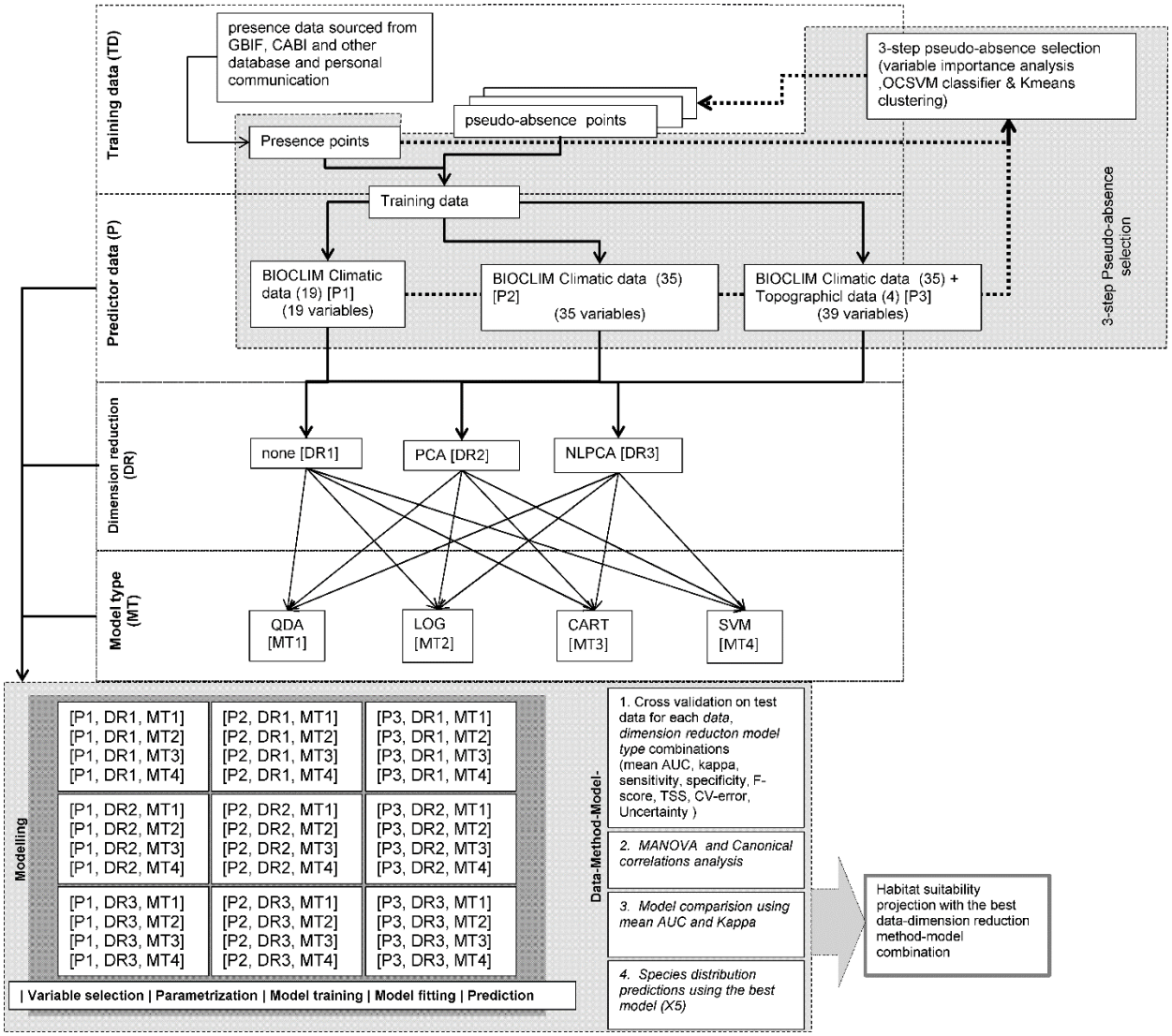


Figure 4.2: Conceptual model showing factorial research design.

To summarise, the whole process produced 180 distribution predictions from 180 models (SP x P x DR x MT combinations). These 180 models were chosen for their average model performance after each model was run with 20 replications of train-test cycles that added up to 3,800 runs considering all the factors and replications.

4.2.2 Predictor data (Abiotic)

Three predictor datasets designated as P1, P2 and P3 were used in this study. The first dataset BIOCLIM19 (P1) consists of 19 variables derived from temperature and precipitation variables of the WORLDCLIM dataset (Hijmans *et al.*, 2005b). This dataset has been used in

many species distribution models and is freely accessible from the open data portal http://www.worldclim.com/bioclim[17].

Table 4.1: Variables included in the three predictor datasets used for in this study.

Variable	Variable Name	Dataset
01	Annual mean temperature (°C)	P1, P2, P3
02	Mean diurnal temperature range (mean(period max-min)) (°C)	P1, P2, P3
03	Isothermality (Bio02 ÷ Bio07)	P1, P2, P3
04	Temperature seasonality (C of V)	P1, P2, P3
05	Max temperature of warmest week (°C)	P1, P2, P3
06	Min temperature of coldest week (°C)	P1, P2, P3
07	Temperature annual range (Bio05-Bio06) (°C)	P1, P2, P3
08	Mean temperature of wettest quarter (°C)	P1, P2, P3
09	Mean temperature of driest quarter (°C)	P1, P2, P3
10	Mean temperature of warmest quarter (°C)	P1, P2, P3
11	Mean temperature of coldest quarter (°C)	P1, P2, P3
12	Annual precipitation (mm)	P1, P2, P3
13	Precipitation of wettest week (mm)	P1, P2, P3
14	Precipitation of driest week (mm)	P1, P2, P3
15	Precipitation seasonality (C of V)	P1, P2, P3
16	Precipitation of wettest quarter (mm)	P1, P2, P3
17	Precipitation of driest quarter (mm)	P1, P2, P3
18	Precipitation of warmest quarter (mm)	P1, P2, P3
19	Precipitation of coldest quarter (mm)	P1, P2, P3
20	Annual mean radiation (W m^{-2})	P2, P3
21	Highest weekly radiation (W m^{-2})	P2, P3
22	Lowest weekly radiation (W m^{-2}	P2, P3
23	Radiation seasonality (C of V)	P2, P3
24	Radiation of wettest quarter (W m^{-2})	P2, P3
25	Radiation of driest quarter (W m^{-2})	P2, P3
26	Radiation of warmest quarter (W m^{-2})	P2, P3
27	Radiation of coldest quarter (W m^{-2})	P2, P3
28	Annual mean moisture index	P2, P3
29	Highest weekly moisture index	P2, P3
30	Lowest weekly moisture index	P2, P3
31	Moisture index seasonality (C of V)	P2, P3
32	Mean moisture index of wettest quarter	P2, P3
33	Mean moisture index of driest quarter	P2, P3
34	Mean moisture index of warmest quarter	P2, P3
35	Mean moisture index of coldest quarter	P2, P3
36	Elevation (m)	P3
37	Slope (deg)	P3
38	Aspect (deg)	P3
39	Hillshade	P3

* *Variables without units are dimensionless indices (Kriticos et al., 2012b)*

The second dataset BIOCLIM35 (P2) includes the BIOCLIM19 variables and additional 16 variables derived from radiation and water-balance soil moisture index (Kriticos *et al.*,

[17] Recent access: 11.02.2014 22:45 New Zealand time

2012b). This dataset is also freely available on the data portal www.climond.org[18]. Detailed information on data construction of these two datasets are given by Hijmans *et al.* (2005a) and Kriticos *et al*. (2012b).

The third dataset, BIOCLIM35+T4 (P3) was prepared by adding an additional set of 4 topographic variables derived from an elevation dataset to the P1 and P2 datasets described above. The four topographic variables consisted of elevation, slope, aspect and hillshade.

The digital elevation model (DEM)[19] data was downloaded from the WORLDCLIM data portal[20] (Hijmans *et al.*, 2005a) and the slope, aspect and hillshade datasets were calculated from the elevation data using ESRI's ArcInfo® spatial analyst software. The principles behind the conversion/ calculation of Slope, Aspect and Hillshade values from the DEM dataset are given in Appendix 4.2.

A 3 X 3 pixel focal area was used in processing all three DEM derived topographical variables. The above three topographical variables and elevation were added to the variables described in dataset P2 to make up the third dataset P3. All the variables in P3 have a resolution of 10 arc minute (18.5 km at equator) same as datasets P1 and P2.

4.2.3 Dimension reduction

Dimension reduction methods are types of data pre-processing that are essential when dealing with large number of variables and/or collinearity. One form of dimension reduction is variable selection and it sometimes is also referred as feature selection. Variable selection involves selecting the most important or useful sets of variables from a multi-variable dataset in order to maximize predictive power (Guyon & Elisseeff, 2003; Dormann *et al.*, 2013). While there are supervised and unsupervised methods of variable selection, it's usually supervised methods that are used in species distribution modelling by utilising species presence/absence points as training data to select variables based on individual importance as well as magnitude of variable interactions. Among the many variable selection (and ranking) methods Pearson's correlation coefficient backed by expert

[18] Recent access: 11.02.2014 22:45 New Zealand time
[19] The DEM data was originally accessed from SRTM and GTOPO30 projects by Hijmans *et al.* (2005a)
[20] http://www.worldclim.com/current

knowledge, decision trees and the Information theoretic criteria are frequently used in species distribution modelling (Dormann *et al.*, 2008; Elith *et al.*, 2013). Variable selection preserves the original units and magnitudes of the variables (unless normalized) that is preferable when understanding the direct effect of each variable is important (Guyon & Elisseeff, 2003). However, there are cases where data space dimension reduction (feature construction) becomes necessary. Space dimension reduction/feature construction replaces a number of given input variables with fewer artificial variables that are constructed out of the original variables. There are two major reasons for such data processing, one is to be able to extract the most information from the input variables without involving prior knowledge about variable importance or collinearity among variables. The other is to increase predictive power by revealing data patterns that are difficult to characterise usually due to multi-collinearity in the original dataset. In case of the latter, supervised reduction methods that construct features that are relevant to the study are used. For example, using variables selected based on presence points of a specific species as a base for dimension reduction, reveals features that explain variance in the presence dataset better than a number of variables unrelated to the study system (Guyon & Elisseeff, 2003).

In this study, two commonly used methods and one new method will be used to investigate the effect of dimension reduction on accuracy of species distribution model results. The first method was to select subsets of the most important variables according to a two-step variable selection that involves random forests and step-wise regression. The second method was to construct principal components through linear transformation of the variables using principal component analysis (PCA). The third method was to produce non-linear transformed artificial components using the hierarchical-bottleneck neural network (also known as auto-associative neural network) algorithm (Scholz & Vigario, 2002) that performs hierarchical non-linear principal component analysis (h-NLPCA).

4.2.3.1 Two-step random forest / stepwise regression (DR1)

The random forest algorithm was selected for the first step of variable selection as it can handle a dataset with large number of variables similar to the ones used in this study. The random forest classifier results in low bias selection by averaging over a large ensemble of

high-variance but low correlation trees (Breiman, 2001; Worner *et al.*, 2010). Random forests have been used both as variable selection as well as a main species distribution model with highly satisfying results (Garzon *et al.*, 2006; Buisson *et al.*, 2010; Kampichler *et al.*, 2010; Lorena *et al.*, 2011; Senay *et al.*, 2013; Worner *et al.*, 2014). Stepwise regression was used to further select the most predictive subset of variables from those selected by the random forest classifier. It is important to mention that there is a caution against using stepwise selection methods (Anderson, 2007). However, its use here is justified because of the insignificant difference reported between a full model space search and the step-wise method in a similar factorial study with even more factors by Dormann *et al.* (2008) and the robustness of the random forest classifier used at the first step. The Akaike's Information Criteria (AIC) was used to rank variable importance in this two-step variable selection method. The second step was added to ensure too many variables are not selected. The two-step approach was used because step-wise regression is shown to be unstable when used with large number of variables, hence applying the first step to prune certain variables of less importance with a random forest classifier which is shown to have low bias in variable selection.

4.2.3.2 Principal component analysis (DR2)

The second dimension reduction method used was a principal component analysis (PCA). PCA is a mathematical method that transforms a set of raw variables into linearly uncorrelated variables by mapping the newly transformed data on artificial orthogonal axes (Pearson, 1901). Each new variable from the transformed data are known as principal components of the data where the first principal component explains most of the variance in the data followed by the second principal component explaining most of the remaining variance provided that this data is orthogonal to the first component. The remaining variance is mapped in the same manner where subsequent principal components explain reduced variance (Legendre & Legendre, 2012). The PCA itself or slightly modified versions have been used in ecological modelling either as a dimension reduction method or as the main species distribution model (Hirzel *et al.*, 2002; Dupin *et al.*, 2011). Although results pertaining to data treated with PCA not being significantly different from data with no

dimension reduction method were reported by Dormann *et al.* (2008), the method is included in this experiment because: 1) it is the most used space dimension reduction method in species distribution model studies when one is used, 2) it is important to replicate the result for different species data as well as more models as recommended by Dormann *et al.* (2008).

4.2.3.3 Non-linear principal component analysis (DR3)

The third dimension reduction method used was a hierarchical non-linear principal component analysis (h-NLPCA)[21], a neural network model developed by Scholz and Vigario (2002). While there are many non-linear PCA methods like symmetrical NLPCA (s-NLPCA) (Kramer, 1991), Principal Curves (Hastie & Stuetzle, 1989), Kernel PCA[22] (Schölkopf *et al.*, 1998) etc. the h-NLPCA was chosen because it is reported to be the true non-linear extension of the linear PCA, as it achieves a hierarchical order of principal components similar to the linear PCA (Gorban, 2007). Generally, the h-NLPCA is both scalable and stable similar to the linear PCA method (Appendix 4.3).

Most of the h-NLPCA parameters are internally computed as they get adjusted throughout the iterative learning. Network weights were initialized at random. The weight decay was set at 0.001; maximum iteration was conditionally set by taking either five times the number of observations or 3000 whichever is the minimum.

[21] http://www.nlpca.org/ recent access: 12/02/2014 17:32 NZ time

[22] Scholz and Vigario (2002) discussed that Kernel PCA is however the most similar to h-NLPCA

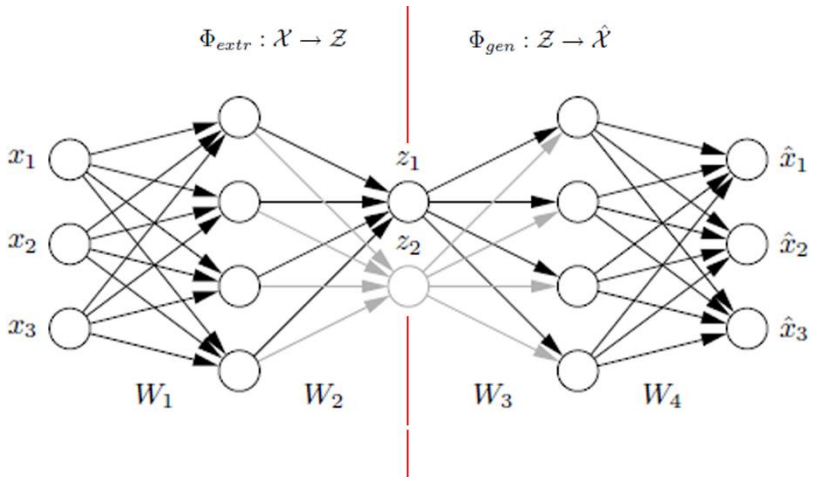


Figure 4.3: The network topology of the hierarchical auto-associative neural network with bottleneck architecture used for the h-NLPCA
the left side of the red line in the middle show the first part of the dimension reduction process where data are extracted non-linearly from the inputs in [x1, x2, x3 …..] and linearly decoded at [z_1, (z_2)] (function Φ_{extr}); the right side of the red line shows where data is linearly decoded from [z_1, (z_2)] and are non-linearly generated at the output [x1, x2, x3…] (function Φ_{gen}). Illustration and description from Scholz et al. (2008, pp. 49-50)

4.2.4 Species data (biotic)

4.2.4.1 Presence data

Five invasive species profiled as highly destructive to ecosystems by the IUCN[23] and the MPI[24] have been used for this study. None of these species have successfully established in New Zealand except for one species whose presence points in New Zealand were kept for validation. However, a number of interceptions of some of the species modelled were reported at New Zealand borders and detections at nearby coastal areas have been recorded.

These insects, which are from a wide range of families, are *Aedes albopictus*, *Anoplopis gracilipes*, *Diabrotica virgifera virgifera*, *Thaumetopoea pityocampa* and *Vespula vulgaris* (established in New Zealand). Ecological background and reviews on the current invasion status of these species is given in Chapter 2. The geographical extents covered by the presence points for these five species widely vary. *A. albopictus* and *V. vulgaris* have a relatively large relative occurrence area (ROA) (Figure 4.4 -A & F) whereas, *D. v. virgifera* and *A. gracilipes* cover an intermediate global extent (Figure 4.4-B & C). *T. pityocampa* (Figure 4.4-E) has the smallest occurrence cover hence smallest ROA with regards to the study area.

[23] IUCN: International Union for Conservation of Nature
[24] MPI: New Zealand Ministry of Primary Industries, Biosecurity department

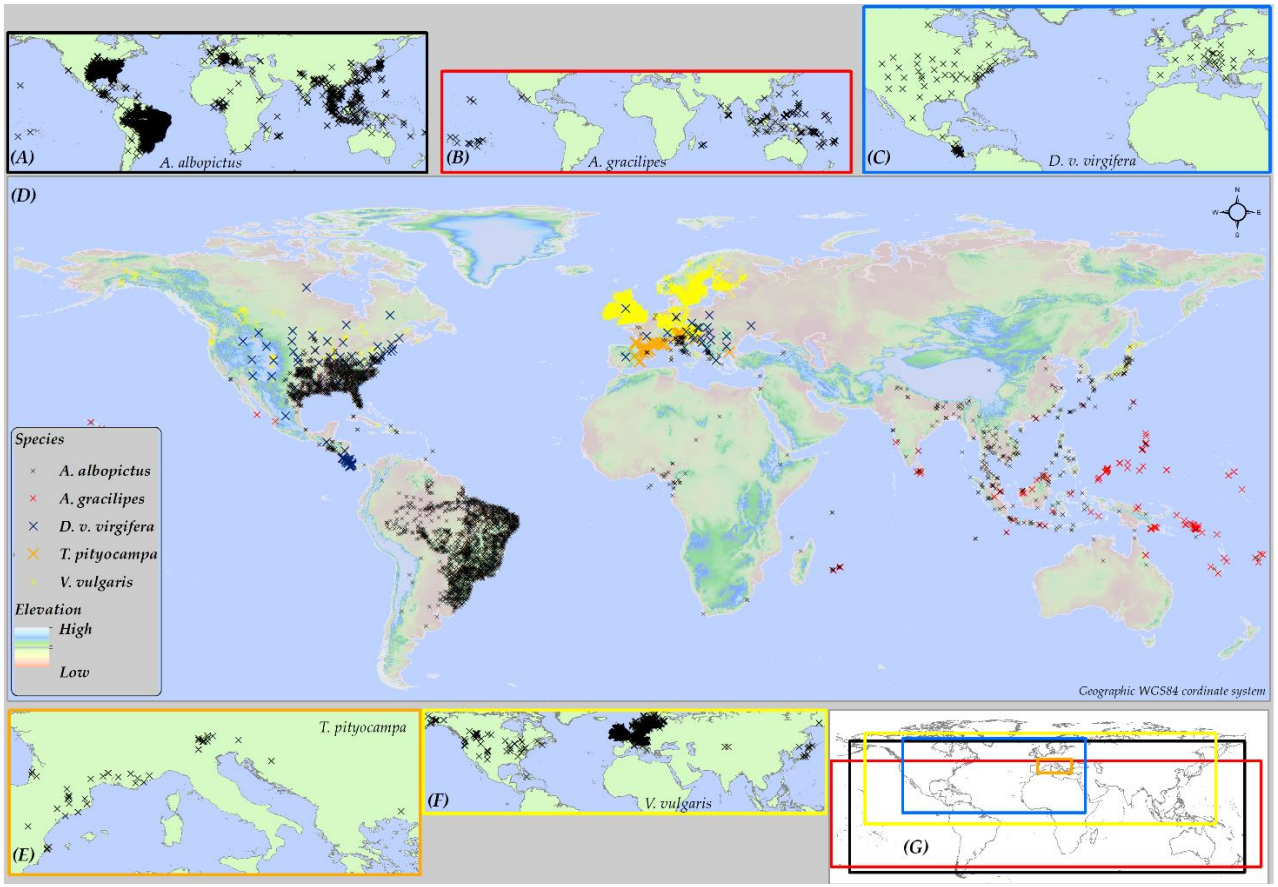


Figure 4.4: Map showing the occurrence data distribution of the five species used in this study. Inset maps show global distribution of (A) Aedes albopictus*; (B)* Anoplopis gracilipes*; (C)* D. v. virgifera*; (E)* Thaumetopoea pityocampa*; (F)* Vespula vulgaris*. (G) Relative global extent of each species; and main map (D) shows the overlaid occurrence distribution of all five species draped on a global elevation model.*

This variation in ROA between the different species tested was needed to effectively investigate the effect of the presence dataset characteristics and their interaction with predictor data and dimension reduction method used on the prediction accuracy of different models. Most of the presence data for the five species was accessed from the GBIF[25] database with some obtained from literature and personal communication with experts.

4.2.4.2 Pseudo-absence data

The 3-step pseudo-absence generation method (Senay *et al.*, 2013) was used to select pseudo-absences for the five species. This method increases model consensus through optimal geographical and environmental discrimination of pseudo-absence points from presence points.

Table 4.2: Number of presence points (available/ spatially unique) and distances used to limit background extent before pseudo-absence selection for the three types of predictor dataset used for the five species in this study.

No.	Species	Predictor	Distance (Km)
1	*Aedes albopictus (3,029/2,928)*	BIOCLIM19	350
2	*Aedes albopictus (3,029/2,928)*	BIOCLIM35	300
3	*Aedes albopictus (3,029/2,928)*	BIOCLIM35+T4	600
4	*Anoplopis gracilipes (385/101)*	BIOCLIM19	550
5	*Anoplopis gracilipes (385/101)*	BIOCLIM35	500
6	*Anoplopis gracilipes (385/101)*	BIOCLIM35+T4	400
7	*Diabrotica v. virgifera (449/84)*	BIOCLIM19	2000
8	*Diabrotica v. virgifera (449/84)*	BIOCLIM35	800
9	*Diabrotica v. virgifera (449/84)*	BIOCLIM35+T4	800
10	*Thaumetopoea pityocampa (67/33)*	BIOCLIM19	300
11	*Thaumetopoea pityocampa (67/33)*	BIOCLIM35	1300
12	*Thaumetopoea pityocampa (67/33)*	BIOCLIM35+T4	800
13	*Vespula vulgaris (10,048/920)*	BIOCLIM19	550
14	*Vespula vulgaris (10,048/920)*	BIOCLIM35	300
15	*Vespula vulgaris (10,048/920)*	BIOCLIM35+T4	700

** Numbers next to species name show available presence points followed by points found spatially unique with regards to the environmental predictor dataset resolution. Another two sets of the above listed datasets were generated according to the above background binding distances for predictor data transformed using PCA and NLPCA making the final number of training/test datasets to 45.*

[25] GBIF: Global Biodiversity Information Facility

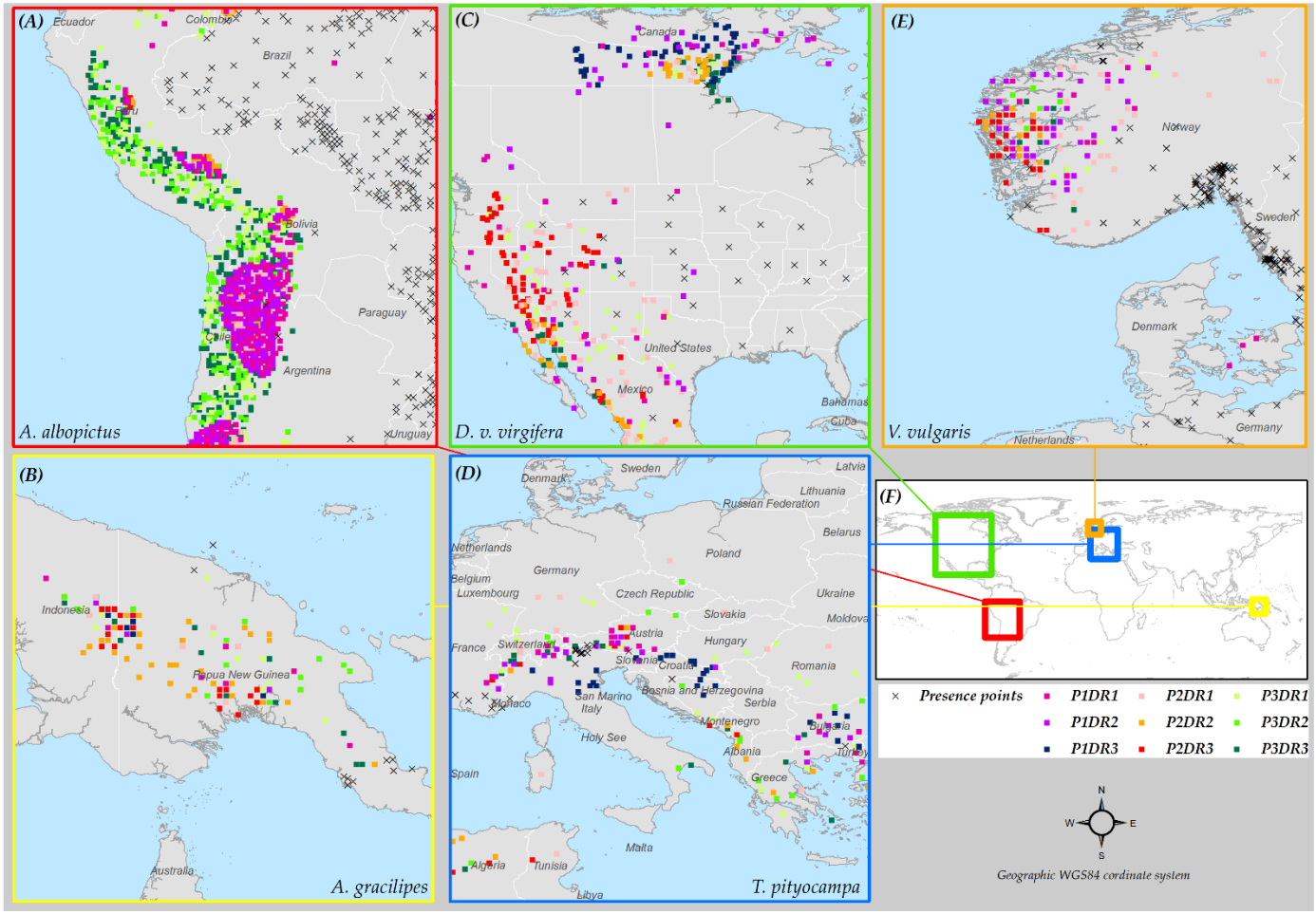


Figure 4.5: Subsets of the global study area with different sets of pseudo-absence points by species, predictor data and dimension reduction method. The nine sets of pseudo-absences generated based on the different combinations of the three predictor dataset and three dimension reduction methods for A. albopictus *(A),* A. gracilipes *(B),* D. v. virgifera *(C),* T. pityocampa *(D) &* V. vulgaris *(E). The extents of the sub-set maps (A-E) are shown on the global map (F).*

One background binding distance was used for each predictor data type, thus a total of three distances were calculated for each species modelled on the three different predictor datasets (P1, P2 & P3) regardless of dimension reduction method used. This is because although dimension reduction methods transform the data into a new configuration they do not add any new data. This distance should be re-calculated every time the presence data or the predictor dataset change because the variable importance analysis over distance is calculated based on a specific correlation structure of variables used. When these variables change the correlation structure also changes making the variable importance over distance pattern different (Senay *et al.*, 2013).

An automated function developed for this study was used to analyse variable importance over distance (VIOD) and generate all the distances at which background data is bound for pseudo-absence selection, the conceptual framework for this custom function is given in Appendix 4.1. Instead of carrying out the entire modelling procedure in Mercator coordinate system (cf-Chapter two), the pseudo-absence points were re-projected to a geographic coordinate system for the reminder of the modelling process, after performing pseudo-absence selection in the equal-grid Mercator Coordinate system. Nine sets of pseudo-absences were generated for each species (Table 4.2).

4.2.5 Relative cover indicators (RCIs)

In this study, indices that were used to estimate how much the training/test data have covered the geographical or the environmental domain of the study area (extent) are termed "Relative cover indicators (RCIs)". The geographical RCI is adopted from the relative occurrence area (ROA) (Jiménez-Valverde *et al.*, 2008). The two environmental RCIs are new indices developed in this study and are described in Section 4.2.5.2 & 4.2.5.3.

4.2.5.1 Relative Occurrence Area (ROA)

Relative occurrence area (ROA) is the ratio of presence data points to the total data points in the study area extent. ROA was first described by Lobo (2008) and later discussed by Jiménez-Valverde *et al.* (2008) to account for the artefacts in modelling species distribution results sourced from the study area extent, which is often subjective and is not related to biological factors. ROA is different from both prevalence and marginality. As described by

Jiménez-Valverde *et al.* (2008) prevalence is the ratio of the number of presences to absences in a given training dataset, and marginality (Hirzel *et al.*, 2002) shows the deviation of the mean environmental condition within the presence data range from the mean environmental condition in the total study area. It is important to consider the ROA of presence datasets especially in factorial studies such as this because it affects model performance (Jiménez-Valverde & Lobo, 2006; Jiménez-Valverde *et al.*, 2008; Lobo *et al.*, 2008; Lobo *et al.*, 2010). In this study no attempt was made to compare results among different species predictions, however the species ROA is assessed to see whether it could explain why different models predict differently for the same species.

4.2.5.2 Environmental relative occurrence ratio (eROR) and environmental relative pseudo-absence ratio (eRAR)

The ROA, calculated in the geographic space, had a limited discriminating power between the difference in the total presence and absence data coverage over varying predictor-dimension reduction method combinations. For example, all the five species had the same ROA over the three datasets as the same global extent was used throughout the study. To discriminate the effect of predictor-dimension reduction method combinations on model accuracy better, two additional simple dimensionless metrics calculated in the environmental space were defined. These metrics were based on the relative area covered by the presence and absence points when superimposed with different predictor-dimension reduction method combinations. All the presence and pseudo-absence datasets of the five species had different corresponding eROR and eRAR values according to the three datasets and the dimension reduction system used prior to their generation. Computing these additional two metrics enabled comparison of the specific effect of dimension reductions on the predictor data in terms of the placement of pseudo-absence points in relation to presence points in the predictor feature space.

The ROA, eROR and eRAR give information on how much of the geographical and/or environmental area is covered by the available training data. The ROA gives the ratio of the presence data with respect to the study area hence can be termed as geographic relative cover indicator, whereas the eROR and eRAR give information about how much of the

environmental space of the total background data is covered by the training/test datasets hence can be referred as environmental relative cover indicators.

How much of the background data is covered by the training and test data could derive differences in predictions of different models for the same species. This is shown in Chapter 3; despite species distribution models being the major source of uncertainty in distribution predictions (Dormann *et al.*, 2008; Buisson *et al.*, 2010), selection of appropriate pseudo-absence points resulted in better agreement between model predictions (Chapter 3). Therefore, even though SDMs are the major source of variation in species distribution predictions, their effect on prediction discrepancies could be reduced by utilizing better datasets that could level the field for SDMs of varying capabilities (Chapter 3).

4.2.5.3 Calculating eROR and eRAR

Step 1: the presence/pseudo-absence training/test data (Table 4.2) points were labelled on each background dataset. For example, for the first BIOCLIM19 (P1) dataset, the presence points from the five species (5 presence datasets) and the pseudo-absence points generated for the five species according to the three dimension reduction methods (15 pseudo-absences), in total points from 20 datasets are geographically overlaid and labelled on points from the P1 dataset. The same procedure was done for P2 and P3 datasets.

Step 2: The three background datasets (P1-P3) on which the presence and pseudo-absence points were labelled, were transformed into a lower dimensional feature space using the standard nonlinear principal component analysis (s-NLPCA)[26] dimension reduction method. This step is done so that the relative covered area by the training points according to all combinations of predictor data-dimension reduction methods can be calculated on a standardized environmental plane. The s-NLPCA (Kramer, 1991) was chosen here to represent each dataset with the best possible low-dimensional components as this method first tests the linear option before proceeding to test non-linear functions (Scholz & Vigario, 2002). The reason s-NLPCA was chosen over the h-NLPCA here was because only

[26] Description of the s-NLPCA is given in Appendix 4.3

dimension reduction and not feature selection was needed here. Two of the principal components with the highest variance were then chosen for each dataset.

Step 3: The three reduced two dimensional feature space datasets from Step 2 were then converted into a raster dataset. Prescribing a user-defined resolution was necessary as the datasets were unit-less because they were all constructed out of proportions of variances from different types of variables with different units (temperature, precipitation, soil moisture, altitude etc.). The default system cell unit in ArcGIS Spatial Analyst was used for the raster conversion. The ArcGIS system definition for the default cell size is given below.

"The default cell size is the shortest of the width or height of the extent of the input feature dataset, in the output spatial reference, divided by 250" (ESRI, 2010)

The default size was chosen because the relative position of the points from one another is the most important characteristics and not the cell size as long as the same resolution was used for the presence and pseudo-absence points within each dataset. Accordingly, resolutions 45, 0.0013 and 0.003 were used for P1, P2 and P3 datasets respectively.

Step 4: Similar to the background datasets, the presence/pseudo-absence datasets for each species were also selected from the two dimensional reduced space datasets using labels applied at Step 1 and converted to raster using the same resolution used for their respective background datasets in Step 3.

Step 5: The environmental relative occurrence ratio (eROR) of each species was calculated by dividing the total area occupied by the presence data of the respective species to the total area of the specific feature space. The environmental relative pseudo-absence ratio (eRAR) was calculated in a similar way by dividing the total area occupied by each type of pseudo-absence for each species by the total relative area of the respective feature spaces (Step 4/Step 3). The relative area occupied by each presence and pseudo-absence data was calculated by multiplying the number of pixels within each data and the respective resolution.

4.2.6 Model types

There are a number of models that can be used to predict species distribution. It is impractical to exhaustively compare all species distribution models, however four methods that more or less represent different modelling techniques were used to illustrate effects of predictor data and dimension reduction and model type choices on species distribution predictions. Even though the sources for each model type used are stated separately in the sub-sections below, the multi-model framework developed by Worner *et al.* (2014)was used to run the four models in a standardized set-up. Moreover, model parameterization and data exporting was also done through this framework.

4.2.6.1 Quadratic discriminant analysis - QDA (MT1)

QDA is one of the classic multivariate models used in species distribution modelling. It is more complex than the basic form linear discriminant analysis (LDA) because it includes quadratic terms in addition to interaction and individual terms (Worner *et al.*, 2010). While QDA makes it easy to assess variable contributions and the prediction in general, it cannot handle dataset where the number of observations are smaller than the variables. Discriminant analysis are commonly used for species distribution modelling studies (Buisson *et al.*, 2010; Kampichler *et al.*, 2010; Worner *et al.*, 2010). The *qda* function from the *R* (R Core Team, 2012) *MASS* (Venables & Ripley, 2002) library was used to run QDA in the multi-model framework.

4.2.6.2 Logistic regression-LOGR (MT2)

Logistic regression model is one of the most frequently used SDMs for species distribution studies (Hartley *et al.*, 2006; Lorena *et al.*, 2011). The *glm* function in the *Stats* (R Core Team, 2012) package in *R* was used to run LOGR in the multi-model framework.

4.2.6.3 Classification and regression trees- CART (MT3)

CART is a classification and regression decision tree that is also frequently used in species distribution models (Garzon *et al.*, 2006; Kampichler *et al.*, 2010). It also has been suggested that decision trees incorporate the complexity needed to explain interactions between high dimensional variable data without a complicated rule that can be easily explained to end-

users (Kampichler *et al.*, 2010) The *rpart* package for *R* was used to run CART in the multi-model framework.

4.2.6.4 Support Vector Machines - SVMs (MT4)

The SVMs are model based on machine learning theory, specifically the artificial neural networks (Vapnik, 1995). SVMs has proven to be an excellent classifier in a number of disciplines; for example in astronomy, medicine, physics, pattern recognition etc. (Way *et al.*, 2012). It has recently been used in ecological modelling along with other machine learning methods like BRT and ANNs (Kampichler *et al.*, 2010; Lorena *et al.*, 2011). SVM here is chosen to represent the complex models used for species distribution in this research framework. Three functions were implemented with the SVM model and these were linear, radial basis and polynomial. The *Kernlab* (Karatzoglou *et al.*, 2004) package for R was used to run SVM in the multi-model framework. A separate optimization of the model has been done in order to specify the best parameters for the different datasets.

4.2.7 Evaluation and validation

4.2.7.1 Multivariate analysis

A multiple factor multivariate analysis of variance (MANOVA) was carried out to investigate the effect of species data, predictor choice, dimension reduction and model types on five dependent variables used to measure model performance (Kappa, AUC, Sensitivity, Specificity and CV error). Precision was dropped from the analysis because the strong positive correlation it had with AUC and Sensitivity. MANOVA was chosen to make an informed decision about which performance measure to use for the evaluation of the factorial design results.

Another reason for the choice of MANOVA was also to increase understanding of the multiple effects of the modelling components (Species data, Predictor choice, Dimension reduction and Model types) on model performance measures derived from two different sources; the first source being the confusion matrix from which Kappa, AUC, Sensitivity and Specificity scores were computed and the second source being the cross-validation error which indicates model consistency predicting the test data correctly within replicates.

Discussion on the appropriateness of MANOVA for understanding the effect of selected factors on a system of multiple responses is given in detail by Enders (2003).

The Mahalanobis' distances derived between each score and the group centroid were compared with the Chi-square ($\chi 2$) distribution and plotted on a q-q plot to determine multivariate normality of the data. The majority of the data conformed to the expected $\chi 2$ value with a few outliers. According to Box's M test, the equality of variance assumption was fulfilled for predictor choice (P) and model type (MT) but not for species data (SP) and dimension reduction (DR). However, I proceeded with the parametric MANOVA test considering the fact that none of the largest standard deviations within a group (factors) were four times larger than the smallest standard deviations within the same group which suggested that MANOVA will still be robust (Howell, 2007).

As a cautionary measure, a conservative α value for the variables that fail the Box's M test was used while carrying out the follow-up post-hoc group comparisons (i.e. α = 0.025 instead of the usual α = 0.05). The effects of predictor data type, dimension reduction and model types on individual model performance indices was analysed using single factor analysis of variance (ANOVA). A non-parametric analysis of variance was done to investigate if change in relative data cover indicators like the ROA, eROR & eRAR has an effect on prediction accuracy. The multivariate statistical analysis was carried out in R (R Core Team, 2012) statistical software version 3.0.2 and with packages agricolae (Mendiburu, 2012), candisc (Friendly & fox, 2013), CCA (González & Déjean, 2012), ggplot2 (Wickham, 2009), heplots (fox *et al.*, 2013), hier.part (Walsh & Nally, 2013) and multcomp (Hothorn *et al.*, 2008).

4.2.7.2 Model performance

The six model performance measures used in this study are listed in Table 4.3. Model Kappa score in combination with cross-validation error were used to identify best and worst data-dimension-reduction-model-type combinations from the 36 scenarios for each species. The model cross-validation error was used to discriminate between models that have comparable Kappa scores. The model with the maximum Kappa score was chosen as the

best model, and the model with the lowest score was considered the worst model. For *V. vulgaris* predictions, external validation was undertaken as additional occurrence data was obtained after the modelling was completed.

Table 4.3: Model performance indices

Index	SDM related review and application	Remark
Kappa	(Allouche *et al.*, 2006)	Developed by Cohen (1960)
AUC	(Jiménez-Valverde, 2012)	Area Under the ROC* Curve
Sensitivity	(Fielding & Bell, 1997; Allouche *et al.*, 2006)	
Specificity	(Fielding & Bell, 1997; Allouche *et al.*, 2006)	
precision	(Worner *et al.*, 2010)	
Cross-validation error	(Worner *et al.*, 2010)	Calculated from cross validation iteration variations

** Receiver operating characteristic curve*

4.2.7.3 Assessing prediction consistency using uncertainty maps

The variability between the predictions across the 36 scenarios for each species were analysed. Mean and standard error maps were also produced so that the spatial pattern of the variability among the models can be easily visualized. The probability density of the standard errors by modelling components (P, DR and MT) were plotted to investigate if any variation found though the multivariate analysis of model performance scores was also reflected in the predicted data.

4.3 Results

4.3.1 Variable selection

Individual variables were ranked based on the frequency of their inclusion in the tested models. The method described by Dormann *et al.* (2008) was adapted for this purpose. The frequency of variable inclusion was calculated based on the number of times a variable was used by a model regardless of whether it was by a straight forward variable selection (RF) or by dimension reduction (PCA, NLPCA).

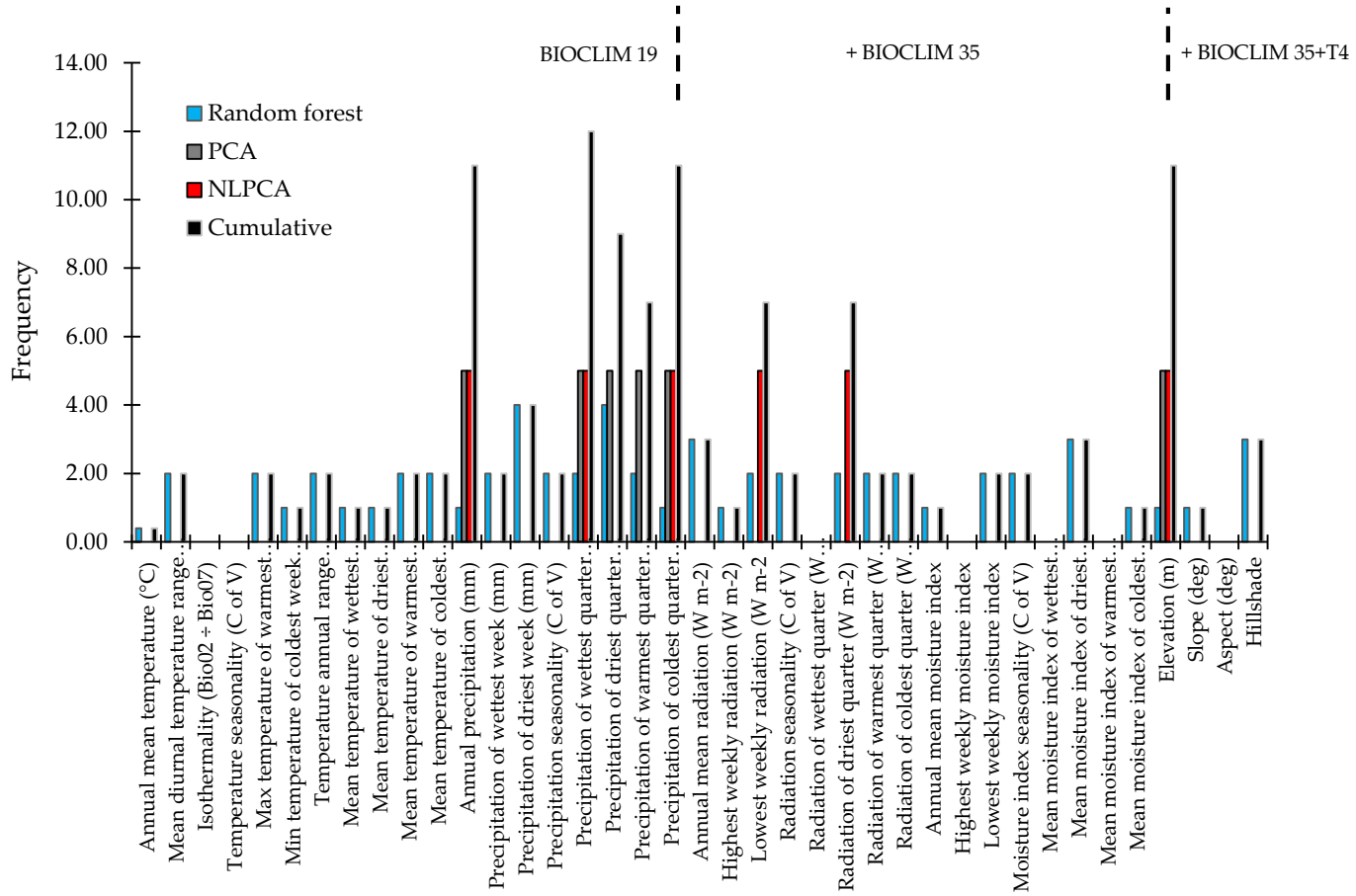


Figure 4.6: Frequency of selection of individual variables across all models
To account for variables that contributed to PCA components the eigenvectors that correspond to principal component scores that explained up to 90% of the variance in the dataset were used (specifically variables with absolute loadings >= 0.32) (Dormann et al., 2013 - & references therein). In the case of NLPCA, due to the non-linear nature of feature extraction it is not possible to get a single corresponding variable coefficient for the scores, however the final weight matrix was used as a proxy for estimating the major contributing variables toward the high variance principal component scores.

The most frequently selected variables (Figure 4.6, Appendix 4.4) were distributed across the three datasets. The first dataset BIOCLIM19 (P1) had temperature and precipitation based variables, while the second BIOCLIM35 [P2] dataset has additional radiation and soil moisture data, and BIOCLIM35 +T4 [P3] had topographic variables additional to those in P1 and P2. The more or less even distribution of frequently selected variables across the three predictor datasets therefore shows that it is important to use a set of predictors that represent various sources of environmental variation to appropriately capture any environmental correlations between a species and its geographical presence.

4.3.2 Multivariate analysis

4.3.2.1 Major trends

The density plot for model AUC, Kappa, sensitivity, specificity, precision and cross validation error are given in Figure 4.7. The plots of Kappa, AUC, Sensitivity, Specificity & Precision scores were skewed to the right showing that the majority of the models performed well above that expected from a random prediction. The density plot of the cross-validation error was skewed to the left showing most of the models had low cross validation error. However, not all of the models that had high AUC scores also had low cross-validation error shown by the CV error density curve being less-leptokurtic than that of the AUC density plots. In this study, the Kappa statistic was found to provide the best discriminatory measure. In Figure 4.7, most of the other scores except cross-validation error are majorly leptokurtic with thin tails showing almost all models performed very well whereas the Kappa plot clearly shows a more spread out distribution implying better model ranking. Further statistical support for the choice of Kappa and CV error as model selection indices is given in section 4.3.2.2.

The MANOVA result (Table 4.4) showed that all the modelling components and the interactions had a significant effect on the linear combination of the five model performance scores with the exception of Predictor choice (P). Predictor choice did not have a significant effect (Pillai's Trace = 0.11, F = 1.50, $\eta^2 = 0.05$). However, it is important to note that the levels in the predictor choice (P) factor were not completely unique as more variables were added from P1 to P2 and on to P3. Change in variables selected as a result of the newly added variables is reported separately in section 4.3.2.1.

A follow up canonical correlation analysis was undertaken and the first canonical variable accounted for 52.4% of the model variance. The corresponding canonical correlation for the first variable was 0.903 (Wilks λ = 0.015, F= 3.53) showing that 81.5% ($0.903^2 * 100$) of the variance in the canonically derived scores was accounted for by the model component factors tested (species type, predictor choice, dimension reduction and model type) (Figure 4.8).

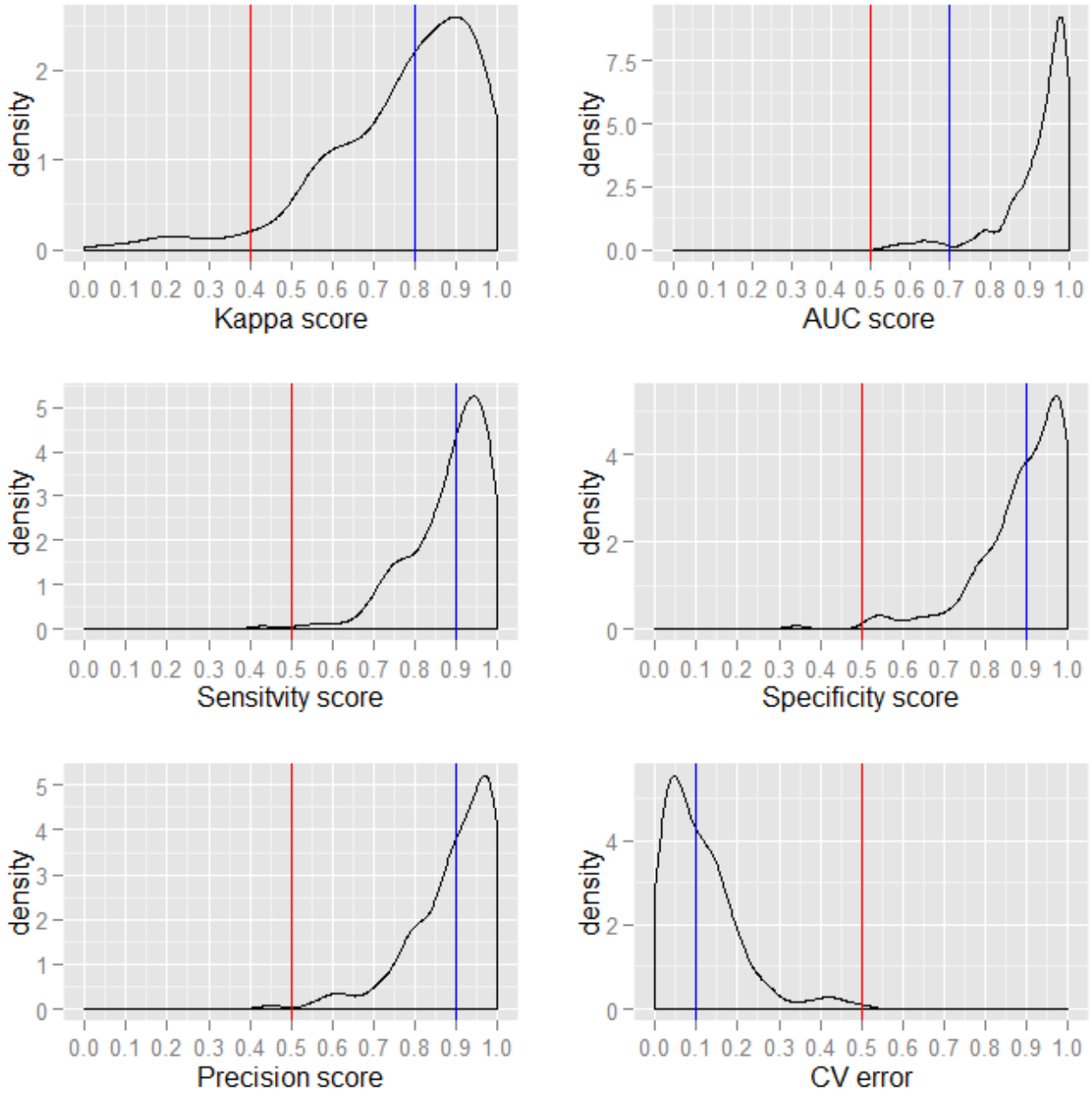


Figure 4.7: Density plot of Kappa, AUC, Sensitivity and Specificity scores for the total 180 models. Red line at 0.5 on the x axis, in cases of AUC, Cross-validation error, Sensitivity and Specificity shows a score expected from a random prediction; and in case of Kappa score indicated at 0.4 on the x axis shows where models are expected to perform worse than a "medium performing model" on the Kappa scale. The blue line show, 0.8 for Kappa, where models are expected to be excellent; 0.7 for AUC a conventional threshold where models are expected to be good; 0.9 for Sensitivity and Specificity, an arbitrarily assigned high performance threshold; and 0.1 for cross-validation error a threshold set as an acceptable training error margin for this study.

Table 4.4: MANOVA results: modelling component effects on model performance.

Modelling components	Pillai's trace	η2.100*	F	Df	p
Model type (MT)	0.79	26.22	9.24	3	<0.0001
Dimension reduction (DR)	0.42	21.01	6.86	2	<0.0001
Species (SP)	0.81	20.32	6.68	4	<0.0001
Predictor (P)	0.11	5.50	1.50	2	0.138
Species x Predictor	0.68	13.51	2.58	8	<0.0001
Species x Dimension reduction	0.58	11.65	2.18	8	<0.0001
Predictor x Dimension reduction	0.49	12.37	3.70	4	<0.0001
Species x Predictor x Dim. Red.	0.95	18.98	1.93	16	<0.0001
Residuals		26.22		132	

** The effect size (eta square) is multiplied by a factor of 100 for easy reporting*

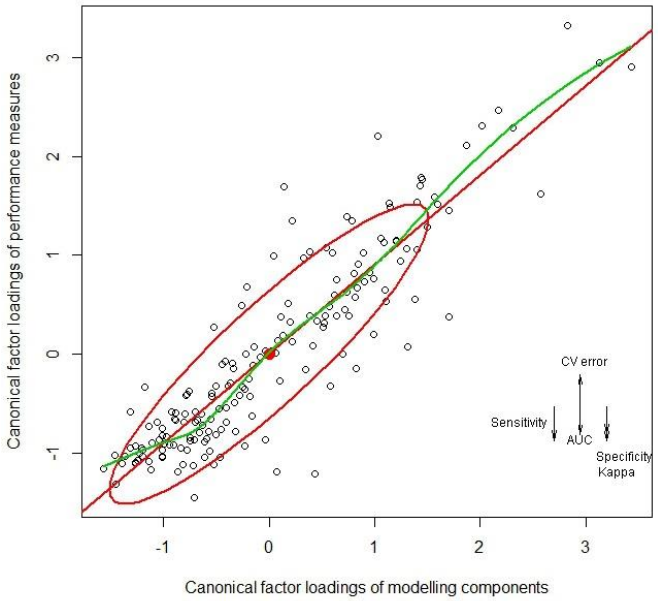


Figure 4.8: Structure correlations (canonical factor loadings) for the first canonical dimension Arrows show the vector direction of variables that correspond to the canonical component on the y-axis. The corresponding variables for the x-axis (combinations of modelling components) were not labelled so not to overcrowd the graph.

4.3.2.2 Model performance measure selection

The canonical correlation analysis was used to determine the model performance measures that most described the effects of the modelling components. The standardized coefficients of the canonical correlation analysis showed that the Kappa score contributed most of the variance of the first canonical variable (79.9%) and cross-validation error contributed the most for the second canonical variable (62.7%). The strong, negative correlation between Kappa and cross-validation error was also a further indication that the multivariate analysis

was supported by appropriate dependent variables as recommended by Tabachnick and Fidell (2001). Therefore, Kappa score and cross-validation error were used to further investigate the significant model component interactions using individual ANOVA and Tukey's Honestly significant difference (Tukey's HSD) post-hoc analysis.

4.3.2.3 Quantifying variance contribution of modelling factors

Individual follow-up ANOVA's were performed for Kappa and cross-validation error scores and the results largely agree with the MANOVA analysis. Even though smaller residuals were obtained for the ANOVA based on cross-validation error scores, the general ANOVA statistic for Kappa and CV error scores were similar. Therefore the statistics for Kappa scores are presented below. All main effects were significant (ANOVA test, SS > 0.24, $\eta 2 > 0.12$, $p < 0.0001$) with the exception of predictor choice (SS = 0.007, $\eta 2 = 0.003$, p = 0.82). All interactions were also significant (ANOVA test, SS between 0.17 – 0.52, $\eta 2$ between 0.09 and 0.22, p between 0.0001 and 0.013).

Hierarchical partitioning (Chevan & Sutherland, 1991; MacNally, 2000) was carried out to quantify the independent contribution of the modelling factors, species data (SP), predictor choice (P), dimension reduction (DR) and model types (MT) on mean Kappa and cross validation scores. Accordingly, species data (SP) was identified as the source of the largest variation both in Kappa scores and model cross-validation errors (54.8% and 47.5 % respectively) followed by model types (MT) which accounted for 38.1% and 43.8% of the variations in Kappa and CV error scores respectively. Dimension reduction (DR) accounted for 6.8% in Kappa score variation and 8.6% in cross validation error variation, and predictor choice (P) scored 0.2% and 0.1% for Kappa and cross validation score variation respectively. The overall trend largely conforms to the results reported by Dormann *et al.* (2008) in their factorial study to quantify modelling uncertainties involving similar modelling components (not including the species data (SP) factor as one species was used in their study). The importance of model types as a source of major variation in predictions is also reported by similar studies (Elith *et al.*, 2006; Pearson *et al.*, 2006; Buisson *et al.*, 2010).

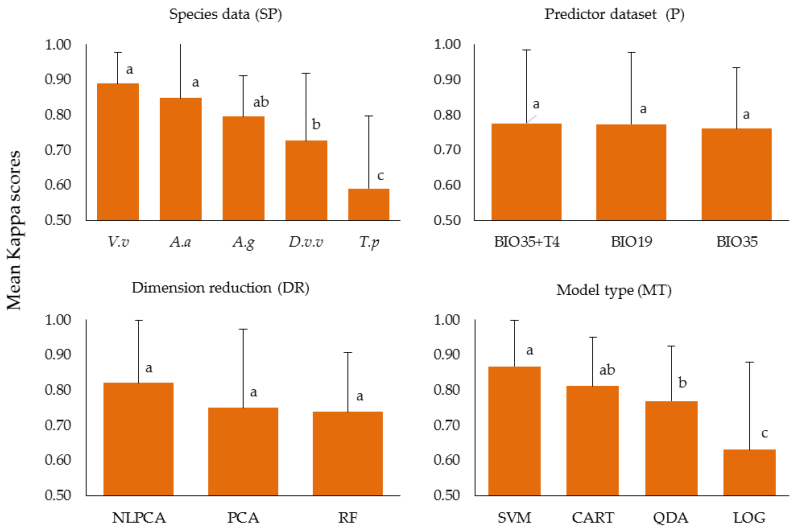


Figure 4.9 Model mean Kappa scores compared over the four modelling components. Error bars indicate the standard deviation over replicate runs. Bars with different letters within a graph indicate statistically significant differences (Tukey's HSD test, α = 0.025 for SP & DR, α = 0.05 for P & MT). Key to factor levels: Species data [SP], A. a = SP1, A. g = SP2, D. v. v = SP3, T. p = SP4, V. v = SP5. Predictor choice [P], BIO35+T4 =P3, BIO19 = P1 and BIO35 = P2. Dimension reduction [DR], RF= DR1, PCA = DR2, NLPCA= DR3. Model type [MT], QDA=MT1, LOG=MT2, CART = MT3, SVM= MT4. The comparison of mean CV errors also showed the same pattern except for a slightly higher CV-error forBIO35+T4 (P3) than BIO19 (P1) which is the opposite of the trend for Kappa scores, however because the differences within the PC group were not significant it was not investigated further.

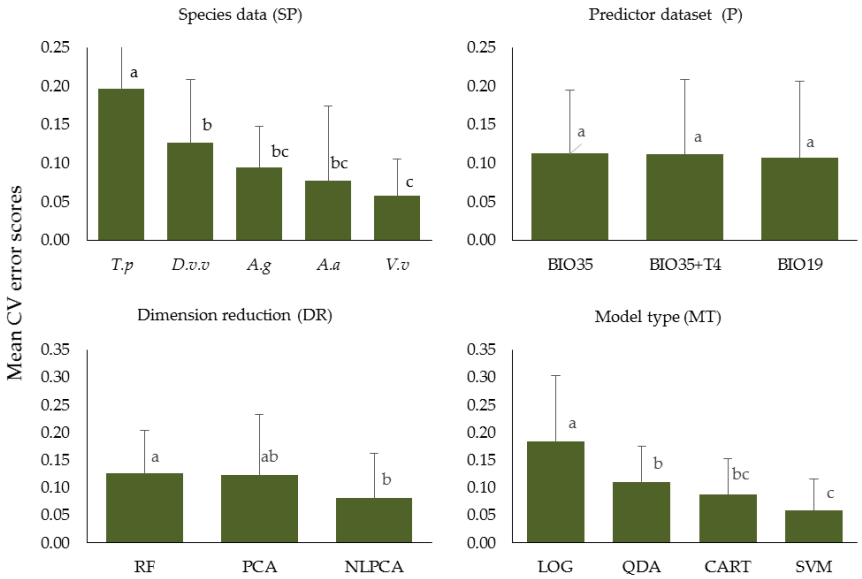


Figure 4.10 Model mean CV error scores compared by the four different modelling components. Error bars indicate the standard deviation over replicate runs. Bars with different letters within a graph indicate statistically significant differences (Tukey's HSD test, α = 0.025 for SP &DR, α = 0.05 for P &MT).

4.3.3 Model components

4.3.3.1 Predictor choice

The multivariate and ANOVA analyses showed that predictor choice (P) did not have any significant effect on model performance scores. However, predictor choice interactions with the other factors were found significant. Although similar results were also reported from other studies, the extremely minimal effect in this study could be because all the variables included in the small predictor dataset (P1) were nested within the bigger dataset levels, P2 and P3, which made assessing effects of individual datasets difficult. The frequency of individual variables selected across all models was assessed to determine the importance of variables (Figure 4.6).

4.3.3.2 Dimension reduction

Dimension reduction had a significant interaction with species data, where its effect on both Kappa and cross-validation error scores varied between species datasets. Non-linear principal component analysis (NLPCA) generally outperformed both linear principal component analysis (PCA) and random forest variable selection method (RF) for all species except for *T. pityocampa* where the score from RF was slightly higher. This is especially true for the two species *A. albopictus* and *V. vulgaris* that had large presence point records that cover large environmental and geographical areas. There was a large difference in Kappa scores due to dimension reduction for some of the species. For example, there was an increase of a magnitude of 0.25 Kappa value for a *D. v. virgifera* distribution model when using NLPCA as opposed to RF. Random Forest (RF) scored better Kappa and low cross-validation error values compared to PCA for *T. pityocampa* and *A. albopictus*, while PCA did better than RF for *D. v. virgifera* and *A. gracilipes*. The mean Kappa scores for RF and PCA were very similar for *V. vulgaris*. The generally poor performance of PCA reported by Dormann *et al.* (2008) was also observed in this study. With regards to interaction with model types there were no clear trends except for the logistic regression model and PCA combinations which consistently gave poorer model performance scores. Based on the results from this study it is not advisable to use PCA with logistic regression species distribution models.

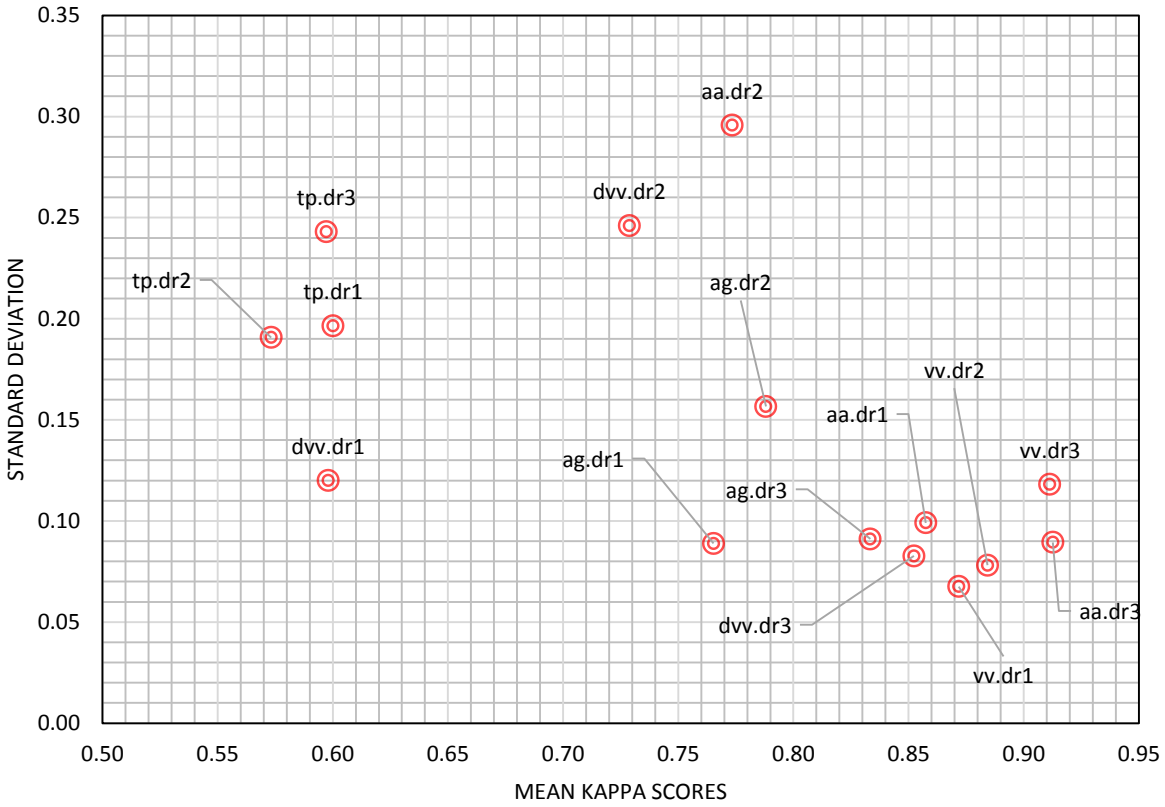


Figure 4.11: A plot of mean Kappa scores against standard deviation over replicates for species – dimension reduction combinations.

4.3.3.3 Model type

The trend of the effect of model type was consistent throughout all combinations of factors. The support vector machine (SVM) model consistently outperformed the other three models. SVM and classification and regression trees (CART) models were consistently ranked with the highest Kappa score and lowest CV-error groups. Logistic regression (LOG) had a generally low Kappa and high CV-error scores throughout the factorial combinations. There is only one instance in which LOG scored better than QDA and CART for *A. gracilipes* within the group of models using the random forest variable selection method.

4.3.3.4 Species data

Species *A. albopictus* and *V. vulgaris* had the highest mean Kappa scores (Figure 4.11), this suggests that the highest prediction accuracy ranks were associated with the species that had the largest presence data records (*A. albopictus*, *V. vulgaris*). Presence data prevalence was consistently associated with high prediction accuracy when compared throughout the

five species. For example, *A. gracilipes* that had higher presence records than *D. v. virgifera* and *T. pityocampa* had higher Kappa scores and lower CV error than both species. However, previous studies showed that prediction accuracy does not necessarily follow size of the presence dataset as also discussed by Elith *et al.* (2006) based on their factorial study involving 226 species and 17 SDMs. Therefore, no conclusion was drawn from the consistency in the relationship between presence data prevalence and high Kappa scores.

4.3.4 Interactions between model components

So far the main effects reported from the multivariate analysis is in accordance with results from previous studies. While there was no significant difference between the dimension reduction methods using mean Kappa values (Figure 4.9), a difference was detected based on the CV error scores of the models (Figure 4.10). This is a good example of how multiple response variables could help investigate possible type II errors in such data analyses where there is limited validation data to repeat the exercise on a separate dataset.

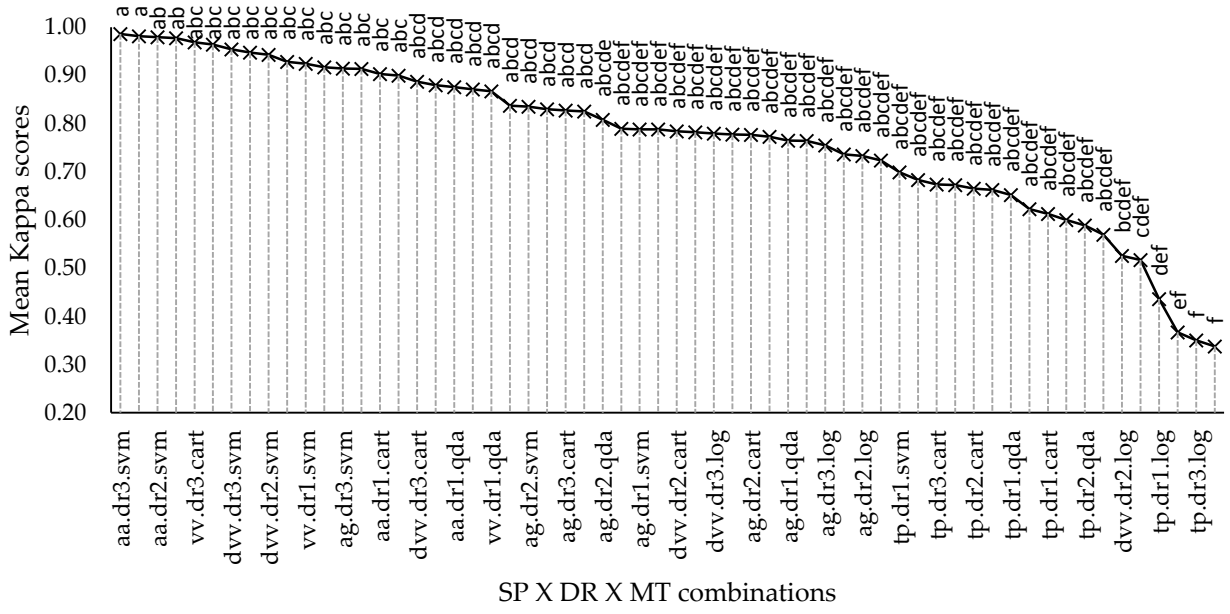


Figure 4.12: Variation in model mean Kappa scores according to different Species data (SP), dimension reduction methods (DR) and model type (MT) combinations. Bars with different letters are significantly different (Tukey's HSD test, HSD = 0.45, α = 0.05).
Only every other data point was marked on the graph (Figure 4.12) to make the x-axis label legible. A table with all the labels is given in Appendix 4.5.

Unfortunately, neither the hierarchical partitioning nor the group comparisons give insight into individual factor interactions. Such, interactions could be essential to determining why different models predict differently for the same species and/or locations. Further investigation was needed to establish if model component interactions have a varying effect on different species data and/or model types. Therefore, the significant two-way and three-way interactions were further analysed using Tukey's HSD test (Figure 4.12).

There were a number of interesting cases where certain combinations did worse, despite belonging to a species with high presence prevalence. For example, the last data point on Figure 4.12 (not labelled) belongs to a prediction made by a PCA dimension reduction method using a logistic regression model for the species *A. albopictus*.

Despite the fact that most of the predictions for *A. albopictus* were ranked high according to Kappa scores (5 out of 9 predictions in the top 10 ranks out of 60 combinations) this particular prediction came last (60^{th}) with a Kappa score of 0.34. This indicates that the PCA dimension reduction method (DR2) was not appropriate for the logistic regression model. Prediction for the same model (logistic regression) and species (*A. albopictus*) scored a Kappa = 0.72 that was ranked at 42 (Appendix 3.7.5) in the comparison when a random forest variable selection was used instead of PCA.

The comparison between species data and model type combinations (Figure 4.13) showed that model type could make a difference to prediction accuracy for some presence data especially when the presence/pseudo-absence data is less reliable. For example, there were no statistically significant differences between Kappa scores for LOG, QDA, CART and SVM predictions for V. *vulgaris*. On the other hand, there was a statistically significant difference between Kappa scores of LOG/QDA and SVM/CART for *T. pityocampa* where the machine learning methods seem to handle the low presence data prevalence better.

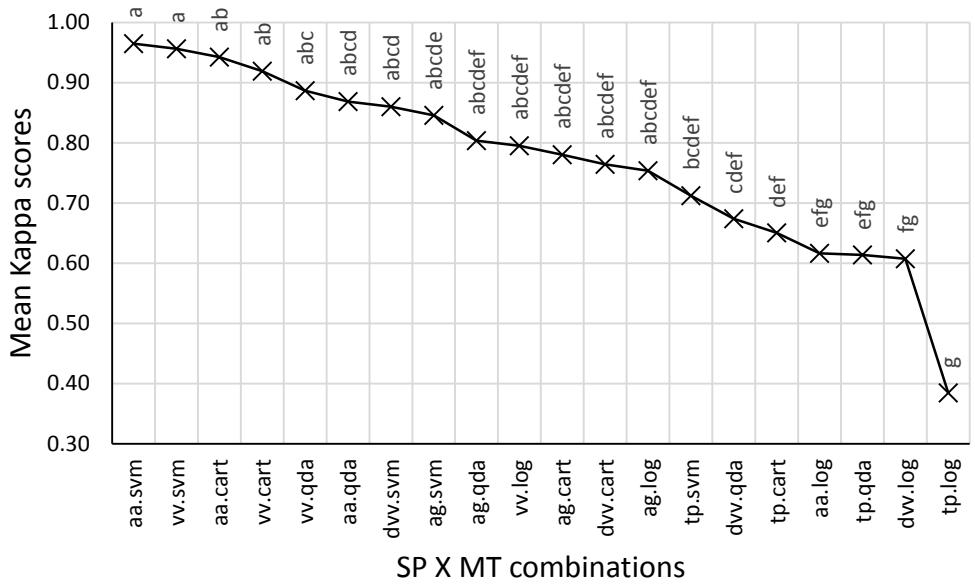


Figure 4.13: Variation in model mean Kappa scores according to different species data (SP) and model type (MT) combinations. Bars with different letters are significantly different (Tukey's HSD test, HSD = 0.24, $\alpha = 0.05$).

4.3.5 Relative cover indicators vs model performance

The relative occurrence area (ROA) values for the presence datasets of the five species in this study are given in Table 4.5. The environmental relative occurrence ratio (eROR) and the environmental relative pseudo-absence ratios (eRAR) calculated for the five species, three predictor and three dimension reduction are given in Table 4.6.

Table 4.5: Relative occurance area (ROA) of the five species

Species (SP)	Area (°²)	ROA.10000*
Aedes albopictus	84.62	1.45
Anoplopis gracilipes	2.92	0.05
Diabrotica virgifera virgifera	2.43	0.04
Thaumetopoea pityocampa	0.95	0.02
Vespula vulgaris	26.59	0.45

** The ROA value here has been multiplied by a factor of 10,000 for ease of reporting. The extremely small ROA values reflect the prevalent problem in global and regional species distribution modelling, as such limited presence records are usually used for prediction over expansive areas. The total area of the global dataset is 16,898.38 °². Area is given in decimal degree square. Each individual presence point is equated to represent the smallest data unit which is the resolution of the dataset (0.17°²)*

Table 4.6: eROR and eRAR ratios for the 45 training datasets

	BIOCLIM19 (P1)				BIOCLIM35(P2)				BIOCLIM35 + T4 (P3)			
	eROR	eRAR RF	eRAR PCA	eRAR NLPCA	eROR	eRAR RF	eRAR PCA	eRAR NLPCA	eROR	eRAR RF	eRAR PCA	eRAR NLPCA
A. albopictus	9.23	11.62	9.97	2.54	**13.37**	10.68	12.53	11.19	11.66	21.53	23.16	**24.08**
A. gracilipes	0.43	0.46	0.46	0.46	0.42	0.64	0.66	0.63	**0.63**	1.01	**1.04**	1.03
D. v. virgifera	0.38	0.39	0.39	0.32	0.60	0.59	0.51	0.54	**0.93**	**0.99**	0.93	0.65
T. pityocampa	0.15	0.15	0.15	0.14	0.25	0.25	0.25	0.23	**0.36**	**0.39**	0.35	0.34
V. vulgaris	3.16	3.99	3.84	2.67	2.77	4.65	4.55	4.46	**4.01**	7.32	**7.93**	5.49

eROR = environmental relative occurrence ratio; RF-eRAR = environmental relative pseudo-absence ratio calculated from pseudo-absences selected from variable space selected by random forest; PCA-eRAR = environmental relative pseudo-absence ratio calculated from absences selected from variable space reduced by linear principal analysis; NLPCA eRAR= environmental relative pseudo-absence ratio calculated from absences selected from variable space reduced by non-linear principal component analysis. Numbers in bold indicate the highest eROR and eRAR values for each species, and underlined values show the highest eRAR within datasets (P1-P3). All values have been multiplied by a factor of 100 for ease of reporting.

Analysis of variance was used to investigate if the magnitude of any of the relative cover indicators have any effect on model Kappa scores. A Kruskal-Wallis non-parametric test was used for this purpose as the homogeneity of variance assumption was not met by the indicators. The ROA, eROR and eRAR were found to have a significant effect on model Kappa and CV error scores (Table 4.7).

Table 4.7: Kruskal-Wallis statistic for mean Kappa score values

RCI*	*df*	$\chi 2$	p
ROA	9	68.99	< 0.0001
eROR	14	74.36	< 0.0001
eRAR	42	102.89	< 0.0001

**Relative cover indicators*

4.3.6 Species level model selection

The species level model performance analysis showed that for *A. albopictus*, dimension reduction (DR) (ANOVA, SS = 0.118, p = 0.004), model types (MT) (SS = 0.689, p = < 0.0001) and their interaction (SS = 0.259, p = 0.001) had a significant effect on model Kappa scores. For *A. gracilipes*, only predictor data had a significant effect on model performance scores (ANOVA, SS = 0.132, p = 0.004). But pairwise comparisons of predictor and dimension reduction combinations (Tukey's test, HSD = 0.24, α = 0.05) showed that PCA based dimension reduction gave the lowest Kappa scores for *A. gracilipes* predictions. For *D. v. virgifera*, results similar to *A. albopictus* were obtained except that the interaction between dimension reduction and model types was not significant. For *T. pityocampa*, predictors (ANOVA, SS = 0.212, p = 0.026) and model types (ANOVA, SS = 0.552, p = 0.001) had a significant effect on model performance. Finally, for *V. vulgaris*, only model type had a significant effect (ANOVA, SS = 0.128, p < 0.0001).

Table 4.8: Best and worst model component combinations for the five species in this study

Species	*Best*	*Kappa*	*CVerror*	*Worst*[#]	*Kappa*	*CVerror*
A. albopictus	P_1DR_3SVM*	0.99	0.006	P_1DR_2LOG	0.14	0.433
A. gracilipes	P_1DR_2QDA*	0.96	0.050	P_2DR_2LOG	0.43	0.292
D. v. virgifera	P_1DR_3SVM*	0.98	0.006	P_2DR_2LOG	0.21	0.344
T. pityocampa	P_2DR_2SVM*	0.88	0.009	P_3DR_3LOG	-0.12	0.498
V. vulgaris	P_1DR_3SVM*	0.99	0.004	P_1DR_3LOG	0.56	0.248
	P_1DR_3CART*	0.99	0.005			

**Combinations are the best based on their high Kappa and low CVerror but are not significantly different from the second best combination. [#]All model combinations identified as "worst" for a species had a significantly lower score than the second worst models. CVerror = cross-validation error.*

Model Kappa scores along with cross-validation error were used to select the best model combination for each species. Accordingly, the best and worst data-dimension-reduction-model-type combinations for each species are given in Table 4.8. Worst prediction in this context does not imply the reported dimension reduction or model types are definitely not suited for the particular species, rather the recommendation is specific to the environmental data, presence records and spatial extent used in this study. Discrimination among models

that have similar Kappa scores was possible because the cross-validation error of models was used as a secondary selection method (Figure 4.14).

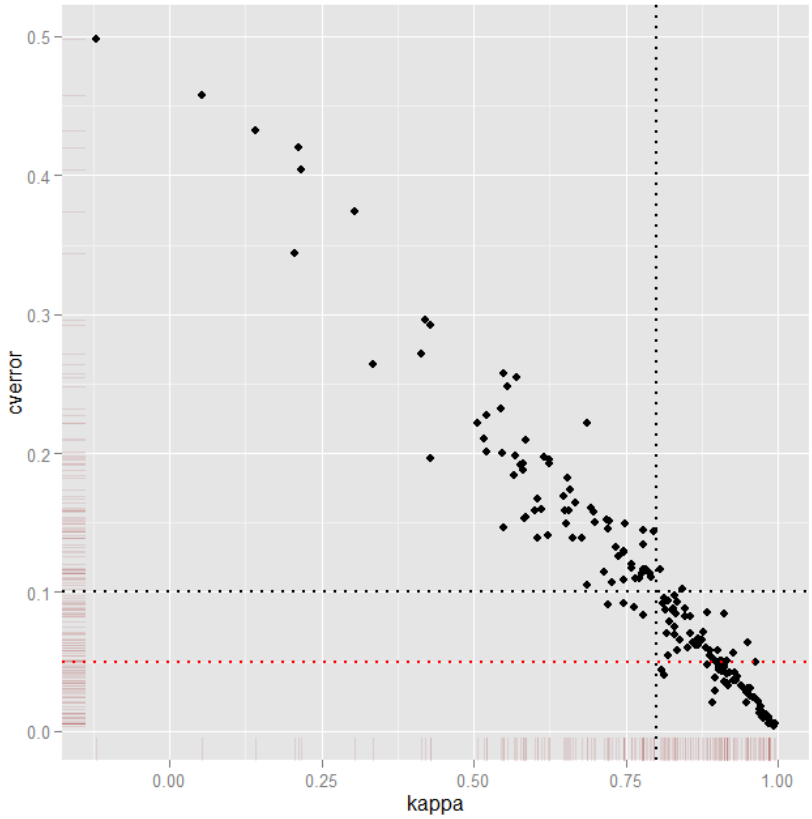


Figure 4.14: Model Kappa scores plotted against cross-validation error scores
Models to the right of the vertical black dotted line have > = 0.8 Kappa score; models below the horizontal black dotted line have < = 0.1 cross validation error and models below the horizontal red dotted line have <= 0.05 cross validation error. The graph shows the advantage of using a second performance score to discriminate between models with similar scores on the first performance measure.

4.3.7 Species distribution predictions and their associated uncertainty

Model predictions were not examined until all the model performance score analyses were finalized. Once the best and worst model combinations for all species were identified, the corresponding predictions were examined. Most of the predictions from the top five best models identified areas that were well described as native or introduced geographical ranges of the five species studied.

However, there were cases where the best model based on the test data was not always the best for prediction. For *A. albopictus* the best model (P_1DR_3SVM) over-predicted with more

than 80% of the global area designated with > 0.9 probability of presence (Figure 4.15–B). Because, the maps were not consulted during the "best model" selection which was based solely on model scores, the subjective bias resulting from selecting the second best model because it seemed more parsimonious (high model performance score with better representation of the prediction in the opinion of the modeller) was avoided.

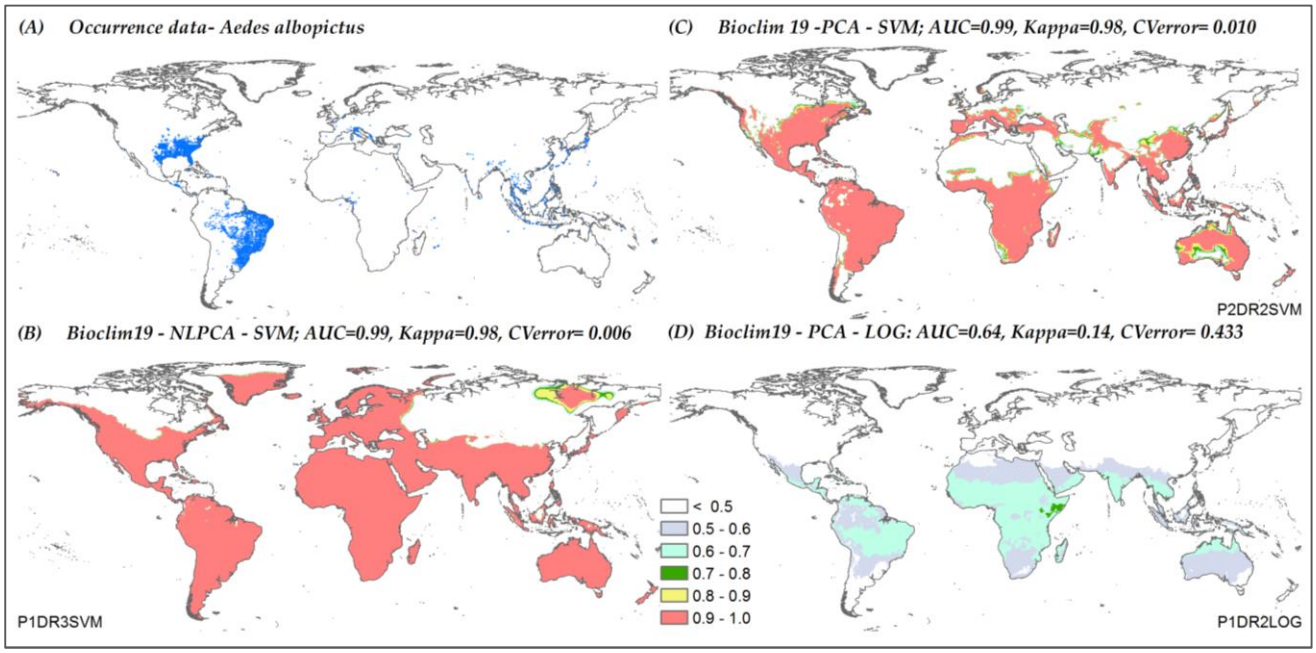


Figure 4.15: Predicted probability of presences for A. albopictus.
(A) Occurrence data, distribution models based on (B) the "best" model (high Kappa and low CV error); (C) the optimum model with similar Kappa and CV error scores as the "best" model but based on validation data with higher eROR and eRAR; (D) the worst model with the lowest Kappa and CV error that also happened to be trained on low eROR and eRAR data.

To avoid introducing subjective bias in model selection, the eROR and eRAR values were consulted to assess the relative coverage of the environmental space by the training/test data using the different data-dimension reduction combinations. It turned out that the model with the best Kappa score and the least cross-validation error for *A. albopictus* prediction also had the lowest eRAR and eROR value for the selected dataset, P1 (Table 4.6, Table 4.8).

As shown in Figure 4.15–B the single "best" model in case of *A. albopictus* over-fitted with extensive areas classified with high probability of presence. The dimension reduction method DR1 (RF) yielded the best eRAR for dataset P1, therefore instead of DR3 (NLPCA)

the dimension reduction method DR1 but with same SVM model and same dataset P1 was chosen as the optimum model for *A. albopictus* (Figure 4.15-C).

Predictions for *D. v. virgifera* (Figure 4.16-B) and *V. vulgaris* (Figure 4.17-B) were similar where the "best" model over fitted. In case of *D. v. virgifera* the "best" model P1DR3SVM had the lowest eRAR for the selected data. But for the same data, dimension reduction methods RF and PCA have higher eRAR. The model P_1DR_2SVM (Figure 4.16-C) was selected as the optimum model for *D. v. virgifera* as the SVM model with DR2 had higher Kappa score than the one with the DR1 method even if they share the same eRAR value.

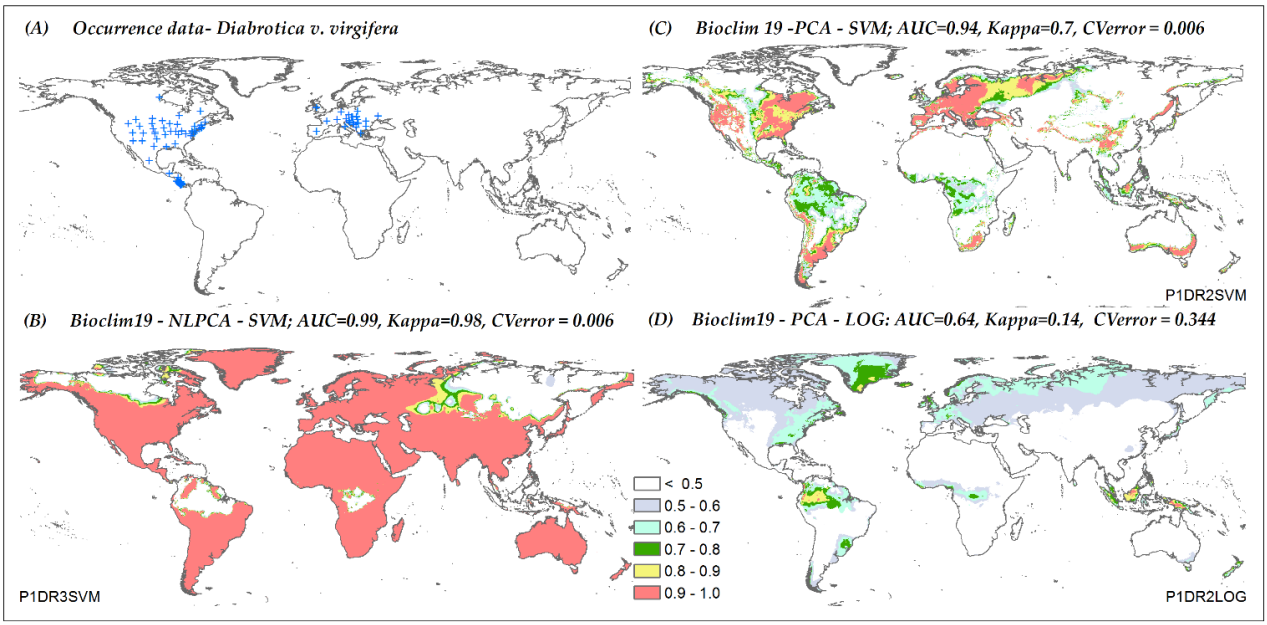


Figure 4.16: Predicted probability of presences for D. v. virgifera.

For *V. vulgaris* the "best" model P_1DR_3SVM covered an extensive global area with predicted presence probability of > 0.9. Similar to *A. albopictus* and *D. v. virgifera* the selected "best" model had the lowest eRAR. Whereas the same dataset and model but with the RF dimension reduction method had higher eRAR. Therefore P_1DR_2SVM was selected as the optimum model for *V. vulgaris* (Figure 4.17-C).

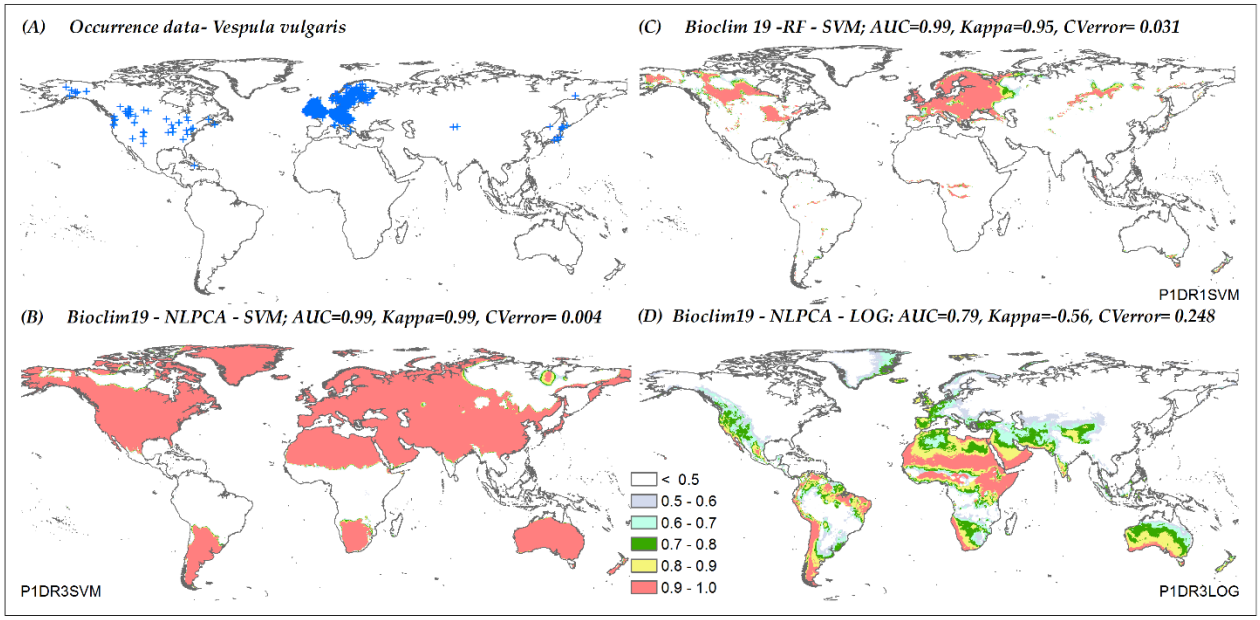


Figure 4.17: Predicted probability of presences for V. vulgaris.

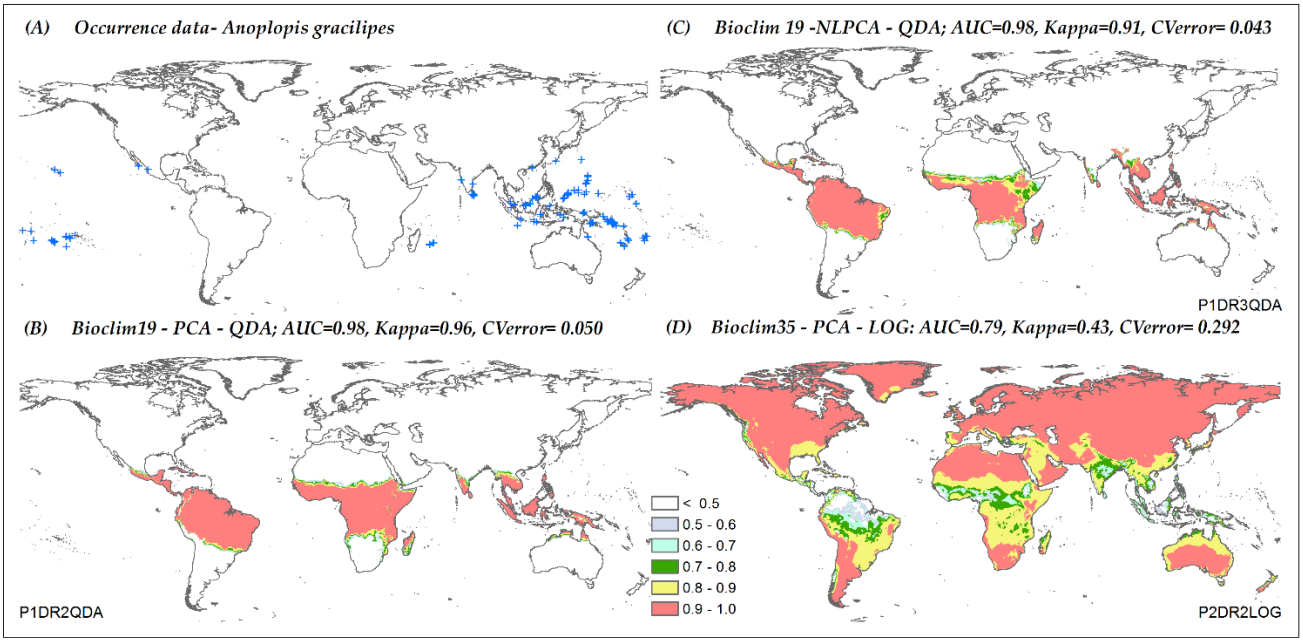


Figure 4.18: Predicted probability of presences for A. gracilipes.

The predictions for *A. gracilipes* (Figure 4.18-B) and *T. pityocampa* (Figure 4.19-B) were not as complicated as the cases discussed above. The "best" model (P_1DR_2QDA) selected for *A. gracilipes* was also trained on the validation data that best represented the environmental

data (high eRAR). In fact, for *A. gracilipes* the eRAR values according to RF, PCA and NLPCA methods were the same for the first predictor dataset P1 (Table 4.6).

In Figure 4.18-C the alternative best model for *A. gracilipes* with the same dataset but a different dimension reduction method and second highest Kappa value was shown. It can clearly be seen that the predictions are similar with the best Kappa model.

The models for *T. pityocampa* have generally lower model performance scores compared to all the other species. The low model performance was expected because of the limited presence data that was available for modelling. The model-data combination P_2DR_2SVM had the best Kappa score and lowest cross-validation error (Figure 4.18-B). The "best" model also had high eRAR for the chosen predictor dataset P2 (Table 4.6). The other combination with high eRAR for the selected predictor dataset as well as high Kappa score was P2DR1 (Figure 4.18-C).

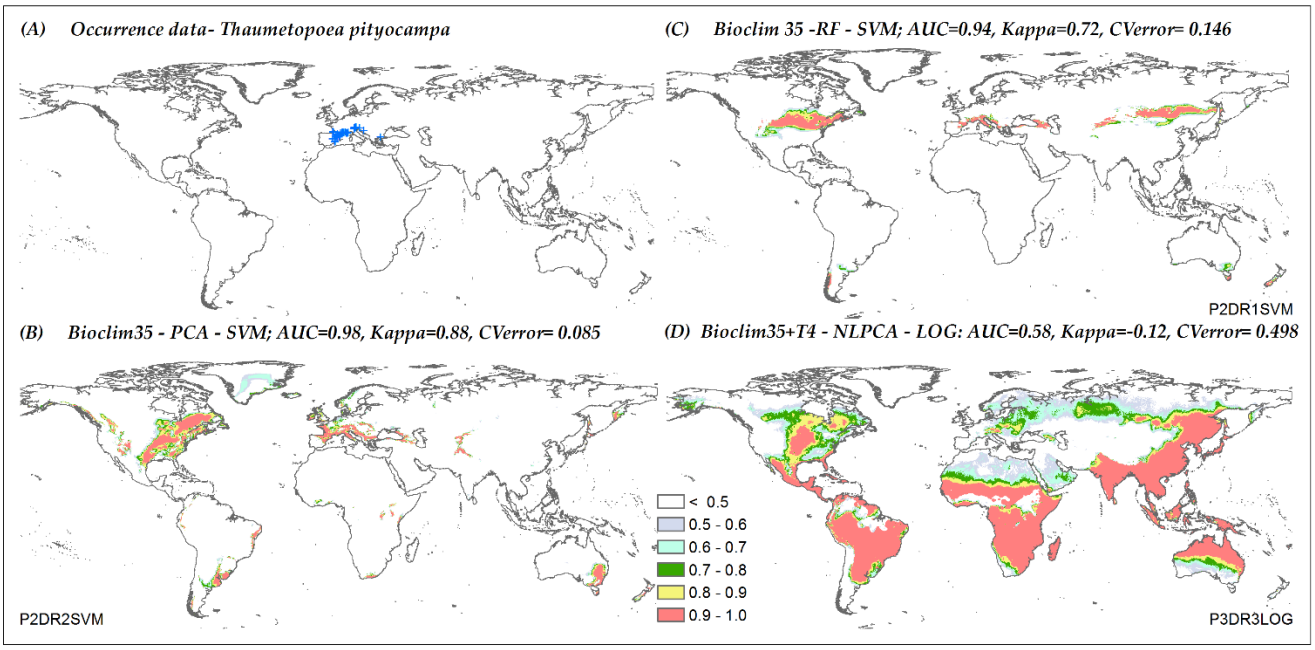


Figure 4.19: Predicted probability of presences for T. pityocampa.

The predictions shown in sub figure (D) for all the five species above (Figure 4.15 -4.19) were the predictions from the worst model-data-dimension reduction method combinations for the respective species. Identifying the worst model was straight forward in all cases because all the model-data combinations identified as the "worst" had a significantly low Kappa

score unlike the "best" models which had the highest Kappa but were not necessarily better than the second or third best models in terms of statistical significance.

The fact that all the "best" three models that over-fitted for *A. albopictus*, *D. v. virgifera* and *V. vulgaris* were based on the same data-dimension reduction-model type combinations (P_1DR_3SVM) called for further investigation. The presence/pseudo-absence data for the three species based on the BIOCLIM19 (P1) dataset and NLPCA (DR3) dimension reduction were plotted in two-dimensional environmental feature space of the P1,P2 and P3 dataset (Figure 4.20) .

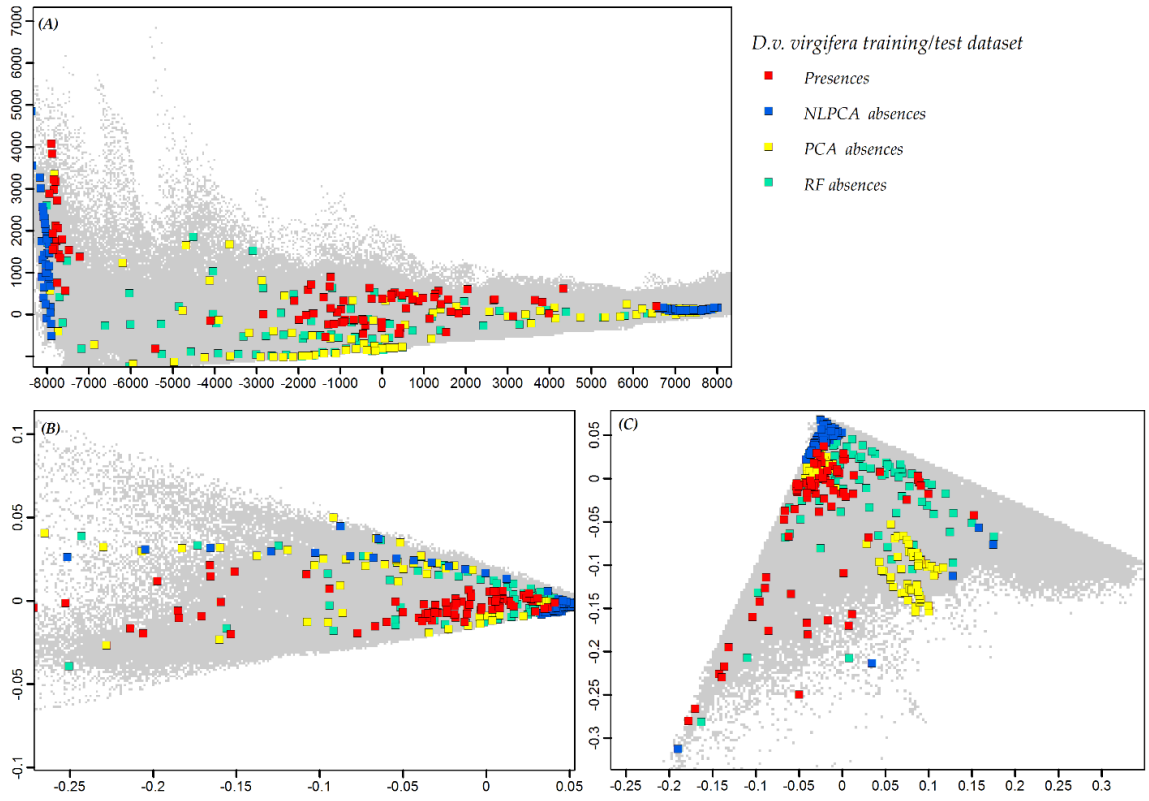


Figure 4.20 Relative locations of D. v. virgifera presences and the three types of pseudo-absences in the environmental spaces of the three predictor datasets.
The training/test data points were plotted on two dimensional environmental space of the three predictor datasets. The axes represent principal components that explained most of the variance in (A) the BIOCLIM19 [P1] dataset (B) the BIOCLIM35 [P2] dataset and (C) the BIOCLIM35+T4 [P3] dataset.

The results showed that the NLPCA pseudo-absences were particularly hyper-discriminated from the presence points for *A. albopictus*, *D. v. virgifera* and *V. vulgaris*, while pseudo-

absences from RF and PCA were still separated but not limited to an extremely limited region of the feature space like the NLPCA pseudo-absences (For example *D. v. virgifera* in Figure 4.20). Such, highly discriminated classification usually leads to over or under prediction (Lobo *et al.*, 2010). However it is highly unlikely in this case due to the use of the 3-step pseudo-absence method that uses a geographical constraint before environmental profiling which ensures even the furthest pseudo-absence point is not extremely separated. The other likely explanation is that models could overfit the highly marginalized pseudo-absence points (low eRAR values) which leads to poor generalization of the unsuitable environmental space leading to over-prediction of suitable areas.

Therefore, I suggest that the most likely reason why the three "best" models selected for their top Kappa score for *A. albopictus*, *D. v. virgifera* and *V. vulgaris* dataset have over-predicted was because they were all trained on the highly discriminated as well as localized NLPCA pseudo-absences.

The variability in model predictions between the 36 scenarios (3P x 3DR x 4MT) for each species was estimated by the standard error across all predictions (Figure 4.21-B for *A. albopictus*). As expected, most of the geographical locations with low variability were close to presence points, however, for all species there were areas with low variability even in areas where there were no presence points in the proximity. Such information gained from the variation in predictions of different models is not as precise as validation data, however it gives more confidence to prediction results.

Even though, only the optimal models and the associated uncertainty maps will be used to discuss the distribution predictions for the species in this study, the average prediction from all the 36 scenarios are also reported. Despite, many warnings in the literature (Kriticos *et al.*, 2013 - & references therein) about averaging predictions from models with different algorithms, the mean prediction maps for the three species with an over-predicting "best" model have noticeably corrected the effect of the over-fitted models (For example, Figure 4.15-B vs Figure 4.21-A for *A. albopictus*).

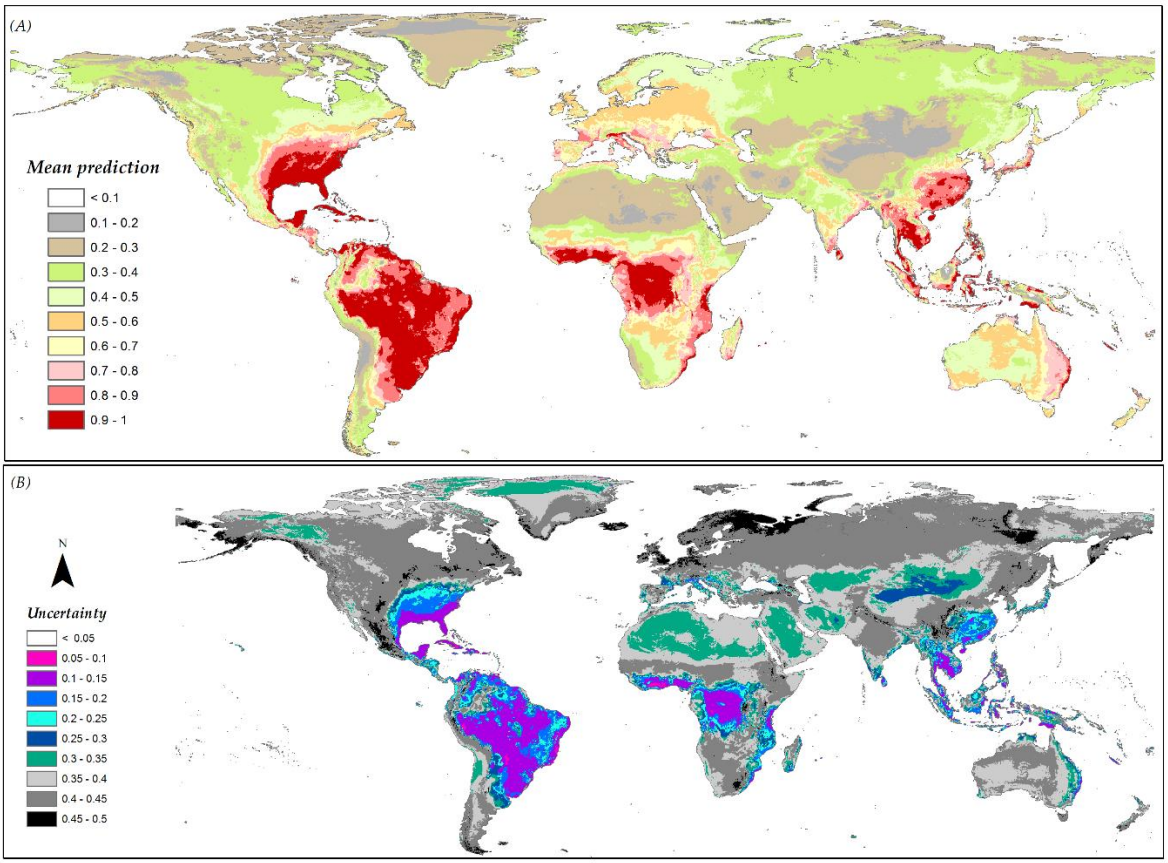


Figure 4.21: (A) Mean predicted presence across all scenarios for A. albopictus*; (B) the associated uncertainty around the mean prediction*

The probability density functions for standard error of model predictions by the different modelling components is given in Figure 4.22-D for *A. albopictus* and in Appendix 4.8 for the other species. Model type had a similar standard error distribution across the five species indicating its constant effect regardless of species data. Dimension reduction and predictor datasets had varied standard error distributions depending on the species. There was considerable spatial variability of predictions according to predictor datasets, even though predictor data (P) did not come out as a major model component in the multivariate model performance analysis. This result shows that model uncertainty analysis should also have a spatial component to determine and understand the full extent of uncertainties in model predictions.

Based on the uncertainty maps the spatial pattern of variability according to model type used (Figure 4.22–C) was not influenced by locations of presence data points as much as the

predictor (Figure 4.22–A) and dimension reduction (Figure 4.22–B) uncertainty maps. This observation, shows that variable prediction power can be gained by using different models even using the same occurrence data. Furthermore, it shows that model type affects spatial characteristics of predictions, which explains the discrepancy in prediction locations among models. The implication of this low spatial auto-correlation between presence locations and magnitude of uncertainty from model types can be important to improve the accuracy of species distribution predictions. For example, if the appropriate model type, given the available species data, environmental data and dimension reduction is selected, improved species distribution prediction could potentially be obtained even for areas that are spatially not close to available presence records. However, more research is needed to confirm this suggestion because the effect of spatial auto-correlation is outside the scope of this thesis.

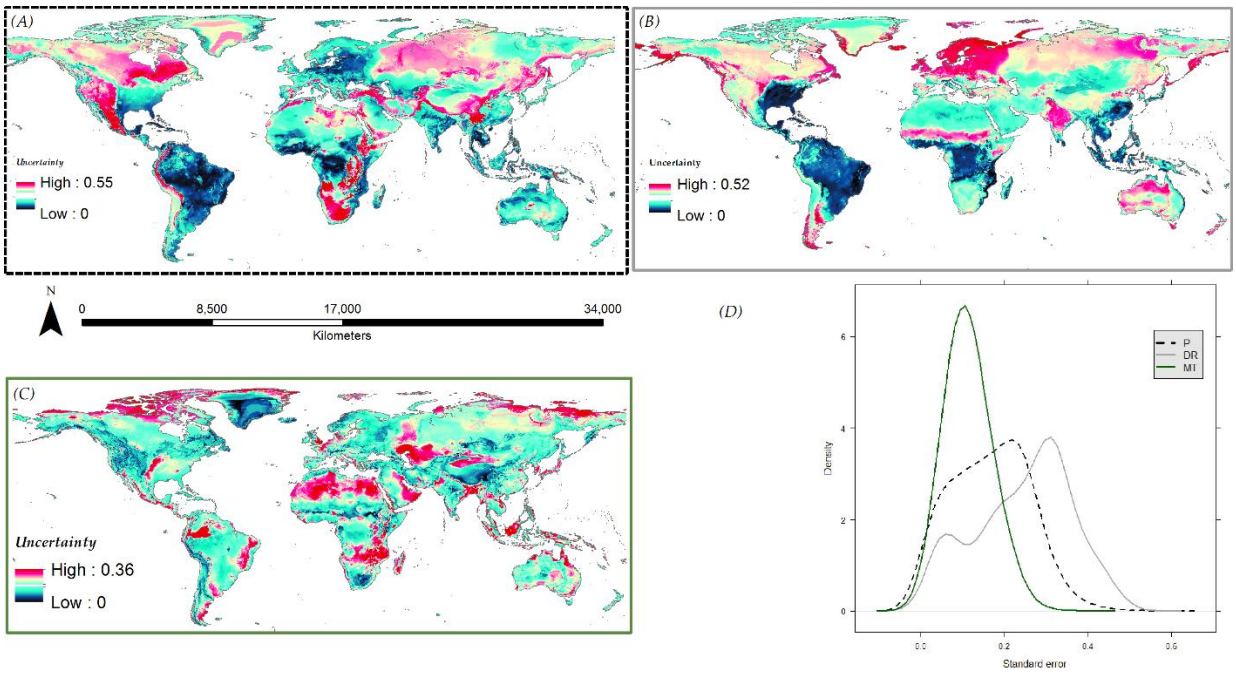


Figure 4.22: Spatial pattern of variability according to (A) Predictor data (P) (B) Dimension reduction (DR), (C) Model type (MT) and (D) the probability density function of the predicted presences by the three modelling components for A. albopictus.

4.4 Discussion

4.4.1 Effects of the major modelling components

As already demonstrated the effects of major modelling components such as model types, dimension reduction, species data and predictor data on model performance is in accordance with previous studies that investigated sources of model uncertainties in SDM

predictions (Elith *et al.*, 2006; Lawler *et al.*, 2006; Dormann *et al.*, 2008; Diniz-Filho *et al.*, 2009; Roura-Pascual *et al.*, 2009; Buisson *et al.*, 2010).

The exception was that the per-species univariate analysis of variance that showed the order of significance of the modelling components was different in one case. For *A. gracilipes* only predictor data had a significant effect on model performance unlike the other four species where model types had the largest effect in accordance with previous studies. Therefore, in such cases where there are exceptions to the established trend further investigation is needed to identify the most important modelling component for further fine tuning of model predictions. For *A. gracilipes* presence locations were clearly clustered in the environmental feature space. All presences were in the Pacific Islands where the data points represented a more or less uniform climate. Such a distinct grouping of occurrence data in the environmental space meant that all models performed well with no significant differences between them.

4.4.1.1 Predictor datasets

Type of predictor dataset used did not have significant effect on model performance in four out of the five species studied. However, the frequency of inclusion of individual variables across the 180 models showed that particular variables from all three predictor datasets were consistently included for all the five species (Figure 4.6 & Appendix 4.4). Because the three predictor datasets were nested within each other and were not independent, it is possible that the effect of some variables was confounded in the multivariate analysis.

According to the analysis of frequency of variables selected more individual variables from the BIOCLIM19 (P1) dataset were consistently included in the models than the BIOCLIM 35 (P2) and BIOCLIM35+T4 (P3) datasets. However, since the P2 dataset contains all the variables in P1, and some variables in P2 were also consistently included in model selection, it is recommended to use the BIOCLIM35 dataset unless the modeller has good evidence the target species distribution can be adequately described by temperature and precipitation derived variables only.

Elevation was the only variable unique to the P3 dataset that was consistently selected. The other three variables slope, aspect and hillshade that were also unique to the P3 dataset were however only included in three models. Therefore, it seems that these three topographical variables may be more useful for higher resolution data at local scales than global or regional scale studies where elevation data is a good proxy for those variables. Despite that, this study was based on five species it may be that further large scale multiple species studies that incorporate these topographical variables may be needed to confirm their redundancy. Alternatively, the variable clustering method proposed by Dormann *et al.* (2013) could be used to check if elevation indeed could be used as a proxy for the other topographical variables for large scale studies.

4.4.1.2 Dimension reduction method

Nonlinear principal component analysis (NLPCA) appeared to perform well based on comparisons of Kappa and cross-validation error. The most important information concerning dimension reduction methods however was obtained from assessing the actual predictions rather than the statistical pairwise comparisons of model Kappa scores.

Three of the five "best" models selected for the five species had used the h-NLPCA as a dimension reduction method and when their corresponding predictions were assessed it was apparent that all the three models had over predicted. Deciding whether model has over-fitted is not an easy task unless there is additional external validation data that covers areas beyond what is covered by the training/test data.

In the case of the three models selected for *A. albopictus*, *D. v. virgifera* and *V. vulgaris* determining whether they were overfitting was straightforward because the areas that were wrongly predicted were in areas where the environmental conditions were outside the known biological tolerance of these three species. For example, some of these areas were Greenland for *A. albopictus*, Sahara and the Middle East for *D. v. virgifera* and *V. vulgaris*. Additional indices such as the relative cover indicators used to assess the best data-dimension reduction combination given the occurrence and predictor data helped to select better data-dimension reduction-model type combinations (see 4.4.2 below). However,

identifying over-prediction is still a difficult problem in cases where over-prediction is not as obvious as the examples given above.

The analysis of the relative locations of the presence points of the three species and their corresponding pseudo-absences points selected from the h-NLPCA (DR3) transformed data, revealed a possible reason for the over-prediction. The pseudo-absences selected from the h-NLPCA transformed datasets were highly discriminated from the presence points as well as localized in the feature space. While high discrimination between presences and pseudo-absences was a welcome attribute of this analysis, the localization of the pseudo-absence points was problematic, leading to less information about the environmental space with a low eRAR.

Therefore, this study suggests that h-NLPCA is not appropriate for presence-absence species distribution modelling that does not involve ensemble prediction or some kind of penalty or regularization to prevent over-fitting of models. However, this method could be an excellent dimension reduction method for presence-only models perhaps in combination with one class SVM (OCSVM) or maximum entropy (MAXENT) (Phillips *et al.*, 2006), and even for the same models used in this study as long as they are modified to incur penalty for over-fitting (Dormann *et al.*, 2013).

The other interesting observation regarding dimension reduction methods in this study was that the only time the h-NLPCA transformed data was selected as optimum was in combination with a QDA model for *A. gracilipes*. This result may suggest that highly localized pseudo-absences in environmental space might affect statistical models like QDA less than machine learning models.

4.4.1.3 Model type

SVM gave consistently the best result for all species except for *A. gracilipes*. But the next best 3-5 models ranked for each species also included CART and QDA and often there was no statistically significant difference between the model that scored the highest Kappa and lowest cross-validation error and the next 3-5 models. That basically means with the appropriate combination of variables and dimension reduction techniques improved

performance of the simpler modelling techniques like the QDA is possible. Logistic regression consistently ranked low especially when used with the PCA dimension reduction method, but all logistic regression predictions with the RF variable selection gave better performance. There are conflicting reports in the literature about using logistic regression in species distribution modelling. But this debate re-enforces the finding in this study that the outcome of each correlative study really depends on the type occurrence data used and data pre-processing methods. What this means could be simply that logistic regression was not the best model type for the five species used in this study given the presence data, the data pre-processing methods and model types used.

Previous studies have shown that regression techniques like GLM and GAM have performed similar to, or sometimes better than machine learning methods (Dormann *et al.*, 2008) and that complex models in general may not be necessarily better than simpler models (Jiménez-Valverde *et al.*, 2008), but the opposite where machine learning methods perform better have also been reported (Elith *et al.*, 2006; Valle *et al.*, 2013).

Based on the results in this study and those in of Chapter 2, model performance depended on data pre-processing including pseudo-absence selection and dimension reduction as well as the training/test data (Lobo, 2008). Machine learning methods were consistently highly ranked for performance for species with high occurrence data prevalence covering large portions of environmental space while QDA did better with species that had low occurrence data prevalence that occupied a localized area in the environmental space. Similar results were reported by Segurado and Araújo (2004) in their study that evaluated commonly used species distributions models. This result leads to the conclusion that each species should be treated individually and model selection should be solely based on the occurrence data to be used and not based on recommendations from other studies that have used totally different presence data and environmental variables. The correlative aspect of the modelling process requires that modelling components and data conditions remain similar for species distribution models even based on the same species to be compared.

More important, most species distribution modelling studies do not specify parameters used in the various machine learning algorithms (But see - Manel *et al.*, 1999; Pearson *et al.*, 2004;

Dormann *et al.*, 2008). With such poor information presentation it is quite possible for end-users to wrongly conclude that machine learning methods such as SVMs and ANNs do not have any advantage. Such a practice is similar to the unacceptability of reporting a statistical analysis without the specific test and assumptions involved.

4.4.2 Prediction evaluation beyond the confusion matrix

Indices that are based on the confusion matrix are the most used method for model performance measurement in species distribution modelling (Fielding & Bell, 1997). These methods were widely and successfully used in other disciplines, especially in clinical studies, long before they were adapted for ecological modelling (McPherson *et al.*, 2004). However, methods based on the confusion matrix are not always sufficient for model validation in the ecological context because of the minimal test data used for predictions of environmental data that cover many orders greater than the test data (McPherson *et al.*, 2004; Lobo *et al.*, 2008; Hanczar *et al.*, 2010). And this test data/ prediction space imbalance is especially pronounced when SDMs are used for regional or global studies.

An example for such a case in this study is when three of the top Kappa score models for three of the five species resulted in an over-predicted distribution. While, the over-prediction in these cases were exacerbated by using the highly localized h-NLPCA pseudo-absences, there is no guarantee that a similar outcome will not occur with other dimension reduction methods or even with expert selected ecologically important variables.

The idea of working with relative cover indicators both in geographic and environmental space was an attempt to include some information on the total background data regardless of whether it was covered by the validation data or not. The method was developed while attempting to use Lobo *et al.* (2010)'s ROA as an indicator of geographic relative cover of species occurrence data with respect to the study area extent. However, the ratios resulting from ROA could not be used to compare different data and dimension reduction combinations, because the ratio does not change for the different data-dimension reduction combinations. Because most species distribution models use the high dimensional environmental space to generate predictions and that the relative cover of validation data in the environmental space varies with the data and dimension reduction method used. It was

necessary to compute environmental relative cover indicators to select the appropriate dimension reduction method with regard to occurrence and background data.

The environmental relative cover indicators eROR and eRAR gave different values depending on the different data-dimension reduction combinations which were used along with model performance scores to select optimum models. For all the five species the optimum model with high model performance score and also the highest eRAR within the selected dataset gave the optimum prediction. Using higher eRAR but lower Kappa scores rather than high Kappa scores with low eRAR values was supported by the pair-wise Tukey's HSD test that showed that the difference in Kappa scores of the top 3-5 models were not significantly different. That makes choosing a model within the same confidence interval as the highest score model but with additional evidence based on better underlying data (in this case the high eRAR), acceptable.

Incorporating cross-validation error along with the Kappa score for model selection was useful in this study because a few models had an almost identical Kappa score (Table 4.8).

4.4.3 Prediction uncertainty and model averaging

Reporting spatially explicit model uncertainty along with species distribution predictions has the advantage of communicating the risk around the prediction. Because of that it has been continually called for in the literature (Guisan & Zimmermann, 2000; Elith *et al.*, 2002). Spatial patterns of low uncertainty in prediction were common in all the standard error maps close to areas where there are occurrences. However, for all species there were areas where low uncertainty was reported even where there were no occurrences near those locations. Such spatially explicit low uncertainty reports enable the identification of hotspot areas where end users could be more certain of the prediction outcome compared with other areas in the study extent.

Standard error maps that reflect geographic variation by the different modelling components used could be very informative for assessing spatial patterns of variability according to the different modelling components. Such visualizations can be used to select specific fitting combinations of data-dimension reduction and model type when conducting

further high resolution studies on subsets of the study extent. The spatial pattern in the standard error maps was informative for assessing effects of modelling components on predictions. Importantly, assessment of the standard error maps and the standard error density plots showed that model type has a uniform spatial effect on predictions as opposed to the other modelling components (Figure 4.22). Additionally, the standard error spatial distribution according to model types was not influenced by the locations of presences. These two observations mean that discrepancy among species distribution predictions mainly occurs due to the difference in ability of models to generalize spatial (Figure 4.22) as well as environmental (Figure 4.20) complexities in the background data.

Model averaging is a controversial subject in species distribution modelling. The major criticism stems from the attempt to average predictions that result from models that have different algorithms (Kriticos *et al.*, 2013). But see Marmion *et al.* (2009)'s justification where they state the variation in algorithms is actually beneficial in terms of the different advantages the different models could offer. Another difficulty is the lack of methods to appropriately rescale probabilistic outputs of different models so that direct averaging of model results can be performed. The mean prediction from the 36 models for each species were reported along with their associated uncertainty map for comparison with the selected optimum model. Interestingly, the mean predictions seem to correct for the models that over-predicted, however mean predictions for the species that had no apparently over-predicting models seem to be more conservative in the spatial coverage of high probability predictions than the selected optimum model. A further discussion of model averaging is beyond the scope of this Chapter but the exercise raised interesting issues. For example, whether pre-selecting only the models that give high Kappa or AUC as suggested by Marmion *et al.* (2009) may not still be optimal when models overfit.

4.4.4 Distribution predictions for the five species in this study

Species level results were briefly discussed in the results section as well as in the previous paragraphs, therefore only observations related with each species and their associated prediction from the optimum model are briefly discussed here.

4.4.4.1 A. albopictus

The data-dimension reduction-model combination (Figure 4.15–C) that comprised the optimum model for *A. albopictus* was BIOCLIM19 with the RF variable selection method and SVM model which was the same combination used for the *A. albopictus* case study in Chapter 2. Therefore, the recommendations given in Chapter 2 regarding the distribution and future research for *A. albopictus* still hold and are not discussed further. Areas identified as having a high climatic suitability for *A. albopictus* could further be assessed by using high resolution data along with trade and cargo network information for the target area because used tyres and plant material imports are identified to be the most important introduction pathways for this species (Scholte & Schaffner, 2007; Scholte *et al.*, 2008).

4.4.4.2 A. gracilipes

For *A. gracilipes*, areas of high probability of predicted presence obtained from the selected P_1DR_2QDA model (Figure 4.18-B) were further assessed by examining the uncertainty map for *A. gracilipes*. The following locations were indicated as highly climatically suitable with overall low uncertainty: Bahi and Amazonas regions of Brazil, the northern coast of Venezuela, Honduras, Nicaragua, coastal areas of Equatorial Guinea, Liberia, Ghana, the western coast of Namibia, south-eastern coast of South Africa (Wetterer, 2005), northern Australia (Hoffmann, 2014) and most islands in the Caribbean and Indian Ocean. *A. gracilipes* is reported to be established in some of the identified areas, even though they were not included in the training or test data as there was no geo-referenced data with the reports. Locations listed with citations are predictions where *A. gracilipes* is already in the country. For New Zealand no high probability areas were predicted. However there was a great deal of variation between *A. gracilipes* distribution predictions for New Zealand by the other models with significantly high Kappa, so maybe this result should be interpreted with caution. On the other hand, *A. gracilipes* have been detected in New Zealand in 2002 in the Auckland area but was later eradicated (Wetterer, 2005). In light of the distribution prediction for *A. gracilipes* in this chapter, it is probable that the success of the eradication could have been enhanced by the unsuitability of the climate in New Zealand.

4.4.4.3 D. v. virgifera

In Chapter 2, it was shown that although the best model for *D. v. virgifera* prediction covered most of the known range of the species, it did not predict the original native range of *D. v. virgifera*, Central America despite presence points from the area being explicitly included in model training. In this study, the optimum model combination P_1DR_2SVM (Figure 4.16-C) predicted the original native range. The only difference between these two models was the dimension reduction method, the same predictor dataset BIOCLIM19 was used in both models. The SVM prediction in Chapter 2 also failed to identify the native range, therefore, model type was not considered a factor. This result implies that models can miss important locations if the appropriate data pre-processing method is not used. Cases where known locations are not predicted even if an occurrence record from the same locality was included in the model training is particularly worrying because modellers will probably not investigate further. Therefore, further study is needed to investigate such scenarios. The recommendations given in Chapter 2 for *D. v. virgifera* are still relevant in this study except that by using the PCA based model in this Chapter it was possible to ensure that the original native range of *D. v. virgifera* was covered and therefore predictions for the rest of the world were conservative. Modelling the climatically suitable areas along with maize plantation cover is recommended to prioritize suitable areas at the risk of *D. v. virgifera* invasion (Aragón *et al.*, 2010).

4.4.4.4 T. pityocampa

The selected model P_2DR_2SVM (Figure 4.20-B) predicted the known geographical ranges of *T. pityocampa* in Europe and Central Asia, except for its distribution in the North Africa (Rousselet *et al.*, 2010). The under prediction was a result of incomplete occurrence data as all the presence points available were from the Mediterranean range of *T. pityocampa* distribution. For New Zealand, most areas in the Manawatu-Wanganui Region, eastern coasts of Canterbury and Otago regions were predicted to be highly climatically suitable for *T. pityocampa* establishment. Because the occurrence data used in this study only covers part of the known ranges of *T. pityocampa*, only positive predictions for the potential distribution of the species were considered. This is essentially because predicted low probability areas

might not necessarily show actual unsuitability due to the missed opportunity of matching areas similar to the North African range for which there were no presence points available to this study. In such cases it is advisable to further consult mechanistic model outputs if physiological information is available for the species (Robinet *et al.*, 2007).

4.4.4.5 V. vulgaris

The SVM model selected for *V. vulgaris* prediction was based on BIOCLIM19 data and a random forest variable selection method. The prediction covered the native Holarctic range of *V. vulgaris* and its introduced range in New Zealand including the Stewart Island and Tasmania in Australia (Thomas *et al.*, 1990; Matthews *et al.*, 2000). An external validation carried out for New Zealand using *V. vulgaris* presence data obtained from the website[27] of Landcare Research Centre showed that 91% of the occurrence sites were correctly predicted by the selected model (Appendix 4.9). Another area identified as a highly suitable was Southern Argentina, *V. vulgaris* was reported from this location in 2010 by Masciocchi *et al.* (2010) but no follow up report on its establishment could be found. However, since the German wasp (*Vespula germanica*) which co-occurs with *V. vulgaris* in New Zealand is present in Argentina (D'Adamo *et al.*, 2002; Lopez-Osorio *et al.*, 2014), it is entirely possible that the climate in the predicted areas of Argentina is also suitable for *V. vulgaris*. If this is the case displacement of *V. germanica* from Argentina is also a possibility according to the trend reported in New Zealand (Harris, 1991). A suitable area of notable size is also predicted in Canada and the U.S.A.

4.4.5 Caveats

1) The geographic and environmental relative cover indicators were useful for understanding the relationship between the training/test dataset and the larger background data into which predictions were made. The eRAR was especially useful for narrowing down the best models along with Kappa and cross-validation error scores. Higher eROR and eRAR values were consistently associated with optimum prediction. However, even if high eRAR values indicated the optimum model for all species in this study there is no strong

[27] http://www.landcareresearch.co.nz/science/plants-animals-fungi/animals/invertebrates/invasive-invertebrates/wasps/distribution/common

evidence that this could not happen by chance. Having an index that can be used to assess model predictions in addition to the Kappa and AUC values to increase the robustness of model selection is very desirable but more research is needed to establish that indices can be used to indicate which predictors and data pre-processing combinations give improved prediction by a given model. Further research may show that the use of such profiling methods could reduce the discrepancy among species distribution models predicting for the same species.

2) In their present form all three relative data cover indicators were calculated based on the single data point they represent, i.e. one occurrence or pseudo-absence data point accounted for one unit of the background dataset. Other methods like minimum convex polygon as used by Elith *et al.* (2006) for the geographical cover and standard deviational ellipses for the environmental relative cover could give less conservative estimate of relative cover ratios.

4.5 Conclusion: why models give different predictions for the same species and locations

Buisson *et al.* (2010) have given a good explanation for why models can predict differently for the same species and study area. Their statement regarding this topic is a very fitting background for this summary and is given below.

"Although SDM are all based on a correlative approach, they use different assumptions, mathematical algorithms, and parameterizations. They may vary in how they model the shape, nature, and complexity of species' response, select predictor variables, weight variable contributions, or allow for interactions." (Buisson et al., 2010, p. 1153)

The investigation of discrepancies between model predictions in this study has shown that it is not appropriate to use such discrepancies as an argument against the robustness of correlative modelling as presented by Kriticos *et al.* (2013). It is apparent that correlative modelling like any scientific technique requires expertise and detailed fine tuning of the methods especially because the correlative approach allows the use of multiple datasets, data pre-processing techniques and model types (Aguirre-Gutiérrez *et al.*, 2013).

Determining why discrepancy among model predictions occurs requires some data mining, visualizing expertise and willingness on the modeller's part, to reveal reasons for discrepancy. As shown in this study the discrepancy is often related to the use of a single

data pre-processing technique for varied model types that have different capacities to model different data well in the interest of keeping a controlled research design.

For example, conventional statistical classifiers cannot handle a dataset where the number of observations is smaller than the number of predictors (Maboudou-Tchao & Agboto, 2013). Various dimension reduction methods are usually used to reduce the number of predictors so that such models could be used. However, doing the same for other machine learning models that can handle large number of predictors just to standardize the study can lead to loss of valuable information that could have been used for improved prediction. That is especially so for a correlative study where the objective is to find predictors that adequately explain the similarity of the presence points while maximizing the difference between the set of presences and absence /pseudo-absences.

Correlative models even when used with the appropriate predictors and data pre-processing methods, may still be affected by incomplete occurrence data from which the potential distribution of species is inferred (Kearney & Porter, 2009; Dormann *et al.*, 2012). That issue is best mitigated as additional occurrence data becomes available. Improving SDMs and associated data pre-processing methods such as pseudo-absence selection, dimension reduction methods and appropriate model specification especially by applying regularization to avoid model over-fitting could lead to a better potential distribution prediction using correlative species distribution models. However, it must be remembered that any improvement will be within the constraints posed by the species environmental range information which could be significantly incomplete.

Chapter 5

5. Incorporating biological traits and environmental adaptation in correlative species distribution models

5.1 Introduction

Species experience novel climate and environmental conditions in the continuum of time. This is especially true for invasive species because the possibility of their contact with new climatic and environmental conditions is much higher due to dispersal into new regions compared to the much slower process of climatic change endemic species experience in their native range (Sutherst, 2000).

Of all species that continuously disperse to new regions due to the increasingly interconnected global transport and trade system, only the ones that arrive in a habitat similar to their native habitat, or species with high tolerance to environmental change manage to establish (Mooney & Cleland, 2001).

Where species are introduced to a habitat similar to their native habitat, it is possible to use species distribution models (SDMs) to understand their potential distribution (Elith & Leathwick, 2009). As a result of increased availability of environmental and presence information on different species as well as continuous research to improve model predictions, use of SDMs has increased both for theoretical exploration of invasion biology and to provide practical tools for surveillance and monitoring of invasive species in applied ecology (Sinclair *et al.*, 2010). When invasive species establish in new environmental

conditions which are not found within their native environmental range, modelling of the potential species distribution is difficult regardless of how SDMs have improved (Diamond *et al.*, 2012).

When there is no biotic and/or physiological information that allows calibrating an invasive species response to the environment, it becomes very difficult to use the preferred mechanistic models (Kearney & Porter, 2009). In this case, correlative species distribution models become an alternative method despite their shortcoming in extrapolating species distribution to environmental ranges outside of the environmental domain of the validation data.

One advantage of a correlative species distribution model over a mechanistic model is the fact that it implicitly accounts for trait variability and evolutionary change of a species (Kearney & Porter, 2009). This is because the modelling process automatically considers the environmental conditions at all presence locations especially if some are outside the known physiological tolerance of the species. This allows to implicitly account for the new environmental ranges species expand through phenotypic plasticity or genetic variation (Helmuth *et al.*, 2005). Nevertheless, correlative species distribution models cannot accurately predict range expansions into environmental conditions that are not included in the model by way of presence points.

In correlative modelling, occurrence data are used to infer the optimal environmental conditions suitable for the species by fitting the model to the values of environmental variables extracted at the occurrence locations (Araújo & Peterson, 2012). This process is highly dependent on the predictors (variables) used to represent or approximate the environmental conditions that are assumed to limit the species distribution (Elith & Leathwick, 2009; Rödder *et al.*, 2009).

Beaumont *et al.* (2009) found models that used presence data from both the invaded and native ranges provided a more complete species distribution prediction than of models that used only native range presence data. Attempting to model the potential distribution of a

species which is still expanding its range using occurrence points only from its native range can seriously underestimate its potential distribution (Rodda *et al.*, 2011).

For correlative models, better prediction accuracy is expected from training data that better represent the realized distribution of the species. However, there are factors that need to be considered while selecting species occurrence points for models. The first and most important is the geographical precision of these occurrence points. Since this is discussed sufficiently in the literature (Rodda *et al.*, 2011) it will not be addressed in this study. The second and less investigated factor is variation between presence points which increases as new presences from invaded ranges are added and its effect on appropriate characterization of the potential distribution of the target species.

Environmental variation between presence points could be a source of uncertainty in correlative model predictions. For example, presence points from the invaded range could be significantly different from the native range due to the species becoming locally adapted to the new habitat. The other notable reason could be a particular population of a species developing biological traits that are distinct from the rest of the population which can allow them to occupy environmentally distinct areas from the commonly occupied range of the species even within the native range.

Such variations within a species often creates multi-modality in environmental data associated with presences. Multi-modality in presence datasets used for species distribution models is rarely discussed in niche modelling studies, but was discussed by Yesson *et al.* (2012) for deep-sea species, which are usually studied at higher taxonomic levels due to lack of species level information. In that study presence points from different species were used in combination to produce habitat suitability at sub-order taxonomic level. Yesson *et al.* (2012) noted that the presence sample distribution was bimodal and that this phenomenon was expected as species are bound to have niche specialization even if they belong in the same sub-order. Hypothetically, local adaption to environmental conditions in a new range could lead to a multimodal distribution within presence data for a single species too. Sangermano and Eastman (2007) showed that seven species distribution models performed

poorly when predicting for a virtual species that had bimodal distribution with regard to the predictors used.

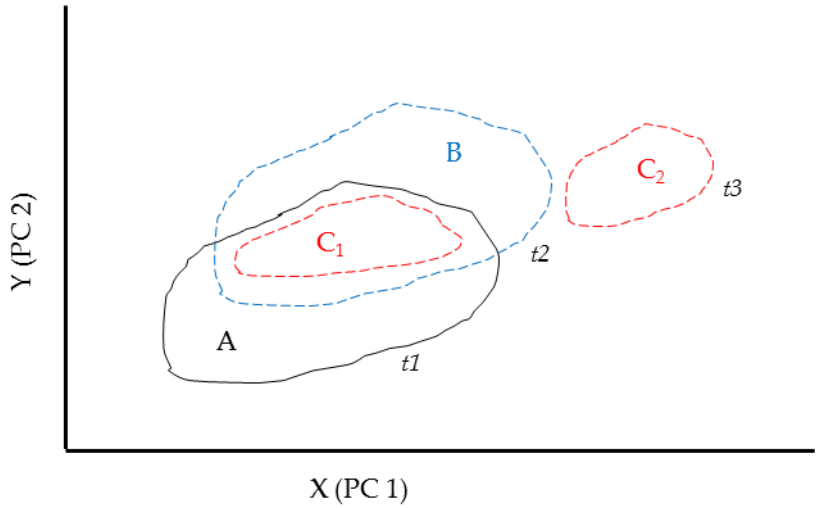


Figure 5.1 An illustration showing the possible effects of variation in environmental ranges obtained from presence data.
Assume that the species under study is an invasive species that is undergoing rapid environmental range expansion through time (t). X and Y denote a lower dimensional environmental feature space (for example, the first two principal components of a PCA transformed set of predictors) constructed out of multiple environmental variables that are selected for species distribution modelling. A, represents the relative position of presence points from the native range in the feature space; B, represents presence points from the invaded range; C, represents a possible alternative environmental adaptation by the species in its invaded range that is non-continuous with the native range unlike the usual circumstance where species are expected to occupy environmental conditions that are contiguous with their native environmental range. In most studies discussed in the paragraphs above the recommendation is to use presences from newly invaded ranges in order to predict the invaded range adequately (using presences from A and B, ensures the prediction of both the native and invaded ranges of the species). However, if the environmental adaptation creates distinct populations (C2) that are adapted to a unique environment compared with the population in the native range (A or C1) then the contribution of populations like C2 can be masked in the overall predictions if models are not parameterized appropriately to handle multi-modality in the presence dataset.

In this chapter, two species were used to investigate the challenges in modelling species with occurrence data of significant variation in terms of response to environmental variables. Occurrence data for western corn rootworm (*D. v. virgifera*) were investigated to see if predictions differed due to significant environmental variation between the invaded and native range of the species. And the great white butterfly (*Pieris brassicae)* occurrences were used to study whether the occurrence of locally adapted population within a species affects SDM predictions. The latter is considered to cover cases where biological traits evolve in response to environmental conditions rather than biotic interactions.

5.1.1 Case study 1: range expansion by **D. v. virgifera**

Three studies that predicted the potential distribution of *Diabrtoica v. virgifera* were used to explore the effect of variation within presence datasets on model predictions. Dupin *et al.* (2011) used training data from both the native and invaded range of the species in North America and Europe. While their model appropriately predicted invaded ranges in these two regions, it did not predict the original range of the species in Central America (Toepfer & Kuhlmann, 2005; Daves *et al.*, 2007). Senay *et al.* (2013) used an improved pseudo-absence method to predict the potential distribution of *D. v. virgifera*. While model Kappa and AUC values were high and the target invaded regions were accurately predicted, the original native region of Central America was not predicted. Aragón *et al.* (2010) used an additional method of climate matching the native range of the species as well as physiological limits to validate the species distribution along with a presence only model (ENFA) to predict suitable areas for *D. v. virgifera*. Their result also shows that the N. American and European ranges of *D. v. virgifera* were appropriately predicted with the exception of the Central American range. The modelling approaches undertaken in these three studies are similar in that they used multiple modelling frameworks and more than one model was used to reach conclusion. *[28]

5.1.2 Case study 2: micro adaptation by **P. brassicae**

Another case which has raised the issue of variation within a presence dataset is modelling the global potential distribution of the great white butterfly (*Pieris brassicae*). While modelling its potential global distribution, it was determined that even if all available global presence points for *P. brassicae* were used (except presence points in New Zealand, which were not accessible at the time) none of the multiple models tested were able to predict the recently invaded area in New Zealand. A literature search to identify more presence data revealed interesting information that a distinct population of *P. brassicae* in southern Spain and Portugal have apparently locally micro-evolved a specific biological trait (Held & Spieth, 1999; Spieth, 2002; Spieth *et al.*, 2011). This species undergoes aestivation (summer

[28] The Central American range of *D. v. virgifera* was predicted using a PCA transformed data in a factorial study reported in Chapter 4. This study investigates the failure of the various models to predict this range using raw variables as most SDM studies do use raw (untransformed) variables.

diapause) in a restricted area to south of the Pyrenees. It is assumed the adaptation is a response to environmental conditions that differ from those experienced by non-aestivating populations of *P. brassicae*, which occur elsewhere in Europe (Spieth *et al.*, 2011).

5.1.3 Research questions

These two cases raise the following questions, were the environmental conditions experienced by the populations of *D. v. virgifera* and *P. brassicae* in central America and southern Spain respectively so different that a single model is not able to fit environmental values associated with these sites and therefore could not accurately predict the potential distribution of these species? In this case do more complex models handle such presence data variation better than simple models? Or maybe a methodology that allows a model to predict distinct populations of a species separately to be combined later, is required?

These considerations led to the hypothesis that using presence data that consists of significant variation within occurrence points that represent different populations might mask the contribution of some of the populations towards the global potential distribution of the species.

5.2 Methods

The seventh objective of the thesis was to investigate the effect of variation in species presence data on model performance and specify methods to detect and handle significant variation in presence datasets.

Two approaches were taken to investigate the effect of variation within presence datasets in this chapter. The first was to investigate the effect of variation between presences from invaded and native range of *D. v. virgifera* and the second was investigating the possibility of distinct environmental adaptation by a population of *P. brassicae* in southern Spain and Portugal. In both cases, distinct presence groups within presence data are modelled separately and their predictions were combined to obtain a final potential distribution map for each species. The combined predictions are compared with the predictions from models that did not consider variation within the presence data.

5.2.1 Predictor dataset, Modelling and Validation (both case studies)

A set of 39 global environmental variables that are derived from temperature, precipitation, radiation, soil moisture and elevation were used for this study (Table 5.1). The first 35 variables were accessed from the CLIMOND website (Kriticos *et al.*, 2012b). The remaining four variables were derived from an SRTM based (NASA-GSFC, 2000) elevation data accessed from the WORLDCLIM data portal (Hijmans *et al.*, 2005b). Details on any pre-processing or derivation performed on/from these datasets were given in Chapter 4.

A high resolution New Zealand extent dataset was prepared by interpolating the above mentioned global dataset into a 30 arc second resolution using a bilinear interpolation method. This was necessary because a high resolution (30') gridded BIOCLIM dataset does not exist for the 35 variables that are provided at 10" resolution on the CLIMOND data portal.

A multi-model framework that includes five models integrated with a pseudo-absence method which uses both geographical and environmental profiling, was used (Worner *et al.*, 2010; Senay *et al.*, 2013). After pseudo-absences were generated variable selection was performed based on the random forest feature selection algorithm using rpart package from R (R Core Team, 2012). Variable selection was done according to the complete presence/pseudo-absence dataset as well as according to the sub-sets of the presence points identified as invaded range and native range for *D. v. virgifera* and aestivating and non-aestivating populations for *P. brassicae*. The procedure followed to identify the independent components in the presence data of *D. v. virgifera* and *P. brassicae* are given in 5.2.2.1 and 5.2.2.2 respectively. The models used for comparison of species distribution predictions were QDA, LOG, CART, SVM and NNET. Kappa scores were used select the best model for Dvv_{I} (*D. v. virgifera* invaded range), Dvv_{N} (*D. v. virgifera* native range), Dvv_{all} (*D. v. virgifera* all presences)[29], Pb_{Aes} (*P. brassicae* aestivating populations), Pb_{NAes} (*P. brassicae* non-aestivating populations), and Pb_{All} (*P. brassicae* all presences) scenarios.

[29] For the Dvv_{all} the result from the *D. v. virgifera* study in Chapter 3 where invaded and native presence are used for the model training is used.

Table 5.1 Variables selected according to the different presence data components

Var. No	Variable Name	DVV_I	DVV_N	DVV_{all}	Pb_{Aes}	Pb_{NAes}	Pb_{all}
01	Annual mean temperature (°C)					✓	✓
02	Mean diurnal temperature range (mean(period max-min)) (°C)					✓	✓
03	Isothermality (Bio02 ÷ Bio07)	✓	✓				
04	Temperature seasonality (C of V)		✓				
05	Max temperature of warmest week (°C)					✓	✓
06	Min temperature of coldest week (°C)					✓	
07	Temperature annual range (Bio05-Bio06) (°C)					✓	✓
08	Mean temperature of wettest quarter (°C)						
09	Mean temperature of driest quarter (°C)					✓	✓
10	Mean temperature of warmest quarter (°C)					✓	✓
11	Mean temperature of coldest quarter (°C)					✓	
12	Annual precipitation (mm)						
13	Precipitation of wettest week (mm)						
14	Precipitation of driest week (mm)			✓		✓	✓
15	Precipitation seasonality (C of V)					✓	
16	Precipitation of wettest quarter (mm)						
17	Precipitation of driest quarter (mm)			✓		✓	✓
18	Precipitation of warmest quarter (mm)						
19	Precipitation of coldest quarter (mm)						
20	Annual mean radiation (W m^{-2})		✓		✓	✓	✓
21	Highest weekly radiation (W m^{-2})					✓	
22	Lowest weekly radiation (W m^{-2}		✓		✓	✓	✓
23	Radiation seasonality (C of V)		✓			✓	✓
24	Radiation of wettest quarter (W m^{-2})						
25	Radiation of driest quarter (W m^{-2})		✓	✓	✓	✓	
26	Radiation of warmest quarter (W m^{-2})					✓	
27	Radiation of coldest quarter (W m^{-2})		✓		✓	✓	✓
28	Annual mean moisture index					✓	
29	Highest weekly moisture index						
30	Lowest weekly moisture index			✓		✓	✓
31	Moisture index seasonality (C of V)			✓			
32	Mean moisture index of wettest quarter						
33	Mean moisture index of driest quarter			✓		✓	✓
34	Mean moisture index of warmest quarter					✓	
35	Mean moisture index of coldest quarter					✓	✓
36	Elevation (m)						
37	Slope (deg)						
38	Aspect (deg)						
39	Hillshade	✓		✓			

**The tick marks show the variables selected for the respective components of* D. v. virgifera *and* P. brassicae *presence datasets as well as the variables selected when the presence datasets are not divided into components.*

5.2.2 Identifying components in occurrence data

5.2.2.1 Cluster analysis – D. v. virgifera

Cluster analysis was used to investigate if there was a significant difference between populations in Central America and the rest of the *D. v. virgifera* range and if it could affect over all model prediction. A presence dataset that incudes geographically referenced *D. v. virgifera* occurrences from N. America, Central America and Europe was prepared. This

dataset comprised 39 environmental variables that were extracted at the recorded occurrence points of *D. v. virgifera.* Principal component analysis (PCA) was used to transform the presence dataset onto artificial orthogonal axes to explain most of the variance in the environmental variables while reducing collinearity.

K-means clustering was performed on the first three principal components of the PCA transformed data. The parameter K for K-means clustering was set to two as the aim was to test for variation between presences from the invaded and native ranges of *D. v. virgifera*. The geographic projection of the clustered presence points showed that all but one of the presence points in Central America were included in one cluster while all the presence points from outside of Central America were included in the second cluster. To denote invaded range the first cluster was labelled I and to denote native range the second cluster was labelled N. It is important to note that *D. v. virgifera* is now considered native to N. America. The reference to the N. American range as invaded here is strictly limited to this study, because here the native range is referenced to Central America due to earlier endemism of the species to that area (Coats *et al.*, 1986).

To test if there was significant variation between these two clusters, the means and standard deviations of the two clusters on the first principal component were assessed.

Let
I = the set of values from the first principal component extracted at the presence points of D. v. virgifera *in the invaded range cluster*
N = the set of values from the first principal component extracted at the presence points of D. v. virgifera *in the native range cluster*

Then the means for each cluster are given by,

$$\bar{I} = \frac{1}{ni}\sum I_j \text{ and } \bar{N} = \frac{1}{nn}\sum N_j \quad \text{eq5.1}$$

Where $I = [I_1, I_2 \ldots I_j]$ *and* $N = [N_1, N_2 \ldots N_j]$ *and ni is the number of presences in the invaded cluster and nn is the number of presences in the native cluster*

$\bar{I}$ and $\bar{N}$ were used to approximate the population mean of the invaded (μ_I) and native (μ_N) ranges respectively.

The standard deviation of the two clusters

$$S_I = \sqrt{\frac{(I_j - \bar{I})^2}{ni-1}} \text{ and } S_N = \sqrt{\frac{(N_j - N)^2}{nn-1}} \quad \text{eq5.2}$$

S_I and S_N were used to approximate the standard deviation of all occurrences in the invaded (σ_I) and native (σ_N) ranges respectively.

The above sample mean and standard deviations of the two populations were used to parameterize a mixed normal random probability density function that contained the sample estimates of the invaded range and the native range as shown in eq. 5.3.

$$f(x|\mu,\sigma) = \frac{1}{2}(f_I(x|\bar{I}, S_I) + f_N(x|\bar{N}, S_N)) \quad \text{eq5.3}$$

The normal probability density function (PDF) of the native and invaded component are given by

$$f_I(x|\bar{I}, S_I) = \frac{1}{\sqrt{2\pi {S_I}^2}} e^{\left(\frac{-(x-\bar{I})^2}{2{S_I}^2}\right)} \text{ and } f_N(x|\bar{N}, S_N) = \frac{1}{\sqrt{2\pi {S_N}^2}} e^{\left(\frac{-(x-\bar{N})^2}{2{S_N}^2}\right)} \quad \text{eq5.4}$$

And the mixture density of the invaded and native range components are given in Eq. 5.5. Proof and justification for the formulae in Eq. 5.4 and Eq. 5.5 are given by Reschenhofer (2001).

$$f_M(x|\mu_I, \mu_N, \sigma_I, \sigma_N) = \frac{1}{2}\left[\frac{1}{\sqrt{2\pi {S_I}^2}} e^{\left(\frac{-(x-\bar{I})^2}{2{S_I}^2}\right)} + \frac{1}{\sqrt{2\pi {S_N}^2}} e^{\left(\frac{-(x-\bar{N})^2}{2{S_N}^2}\right)}\right] \quad \text{eq5.5}$$

The combined normal distribution from Eq. 5.5 was plotted for visual investigation of variation between the two components of the combined dataset.

A likelihood ratio bimodality test was also performed on the complete *D. v. virgifera* presence data, this method is more robust than by simply plotting the mixed PDF of the two samples as it compares the sample distribution against a unimodal curve option and an unrestricted fit set by the sample parameters (Holzmann & Vollmer, 2008). The test confirmed the variation between the two clusters of *D. v. virgifera* presence points, Dvv_I and Dvv_N (Results section 5.3.1).

5.2.2.2 Biological traits as a precursor to environmental variation in presence data - P. brassicae

The *P. brassicae* training dataset was classified into two user defined classes to represent the aestivating and non-aestivating populations of *P. brassicae*. This classification followed the geographic boundaries of the aestivating *P. brassicae* population as per the description by Held and Spieth (1999) and Spieth *et al.* (2011).

The more or less permanent geographical cline (Figure 5.2) that was reported by Spieth *et al.* (2011) to represent the transition between aestivating and non-aestivating populations of *P. brassicae* was constructed using spatial markers given in their publication. All presence points south of the cline in continental Europe were recorded as aestivating and all other presence points were recorded as non-aestivating.

Out of the total 2,241 spatially unique *P. brassicae* presence points, 35 fell into the aestivating class and the remaining 2,206 points were classed as non-aestivating. Assessing multimodality of the *P. brassicae* presence dataset using the equations described above is difficult as the aestivating class represents only 1.5 % of the sample dataset and any significant difference could be due to spurious variation that shows a local maxima due to lack of data. Moreover, environmental variation may remain undetected due to the small number of observations for the aestivating class. This is because datasets with possibly two components do not always need to be bimodal, as well, a unimodal dataset could appear to have two modes if there is no sufficient data to characterise its true distribution (Holzmann & Vollmer, 2008). Therefore, a separate method was employed to check if the contribution of the aestivating population towards the overall potential distribution of *P. brassicae* was masked when using all presence points in model predictions.

The number and type of variables selected according to presence locations from the aestivating and non-aestivating populations were compared. To check for variation between environments associated with aestivating and non-aestivating presence points, their relative position in the feature space of variables selected according to the aestivating presences as well as the non-aestivating presences were mapped.

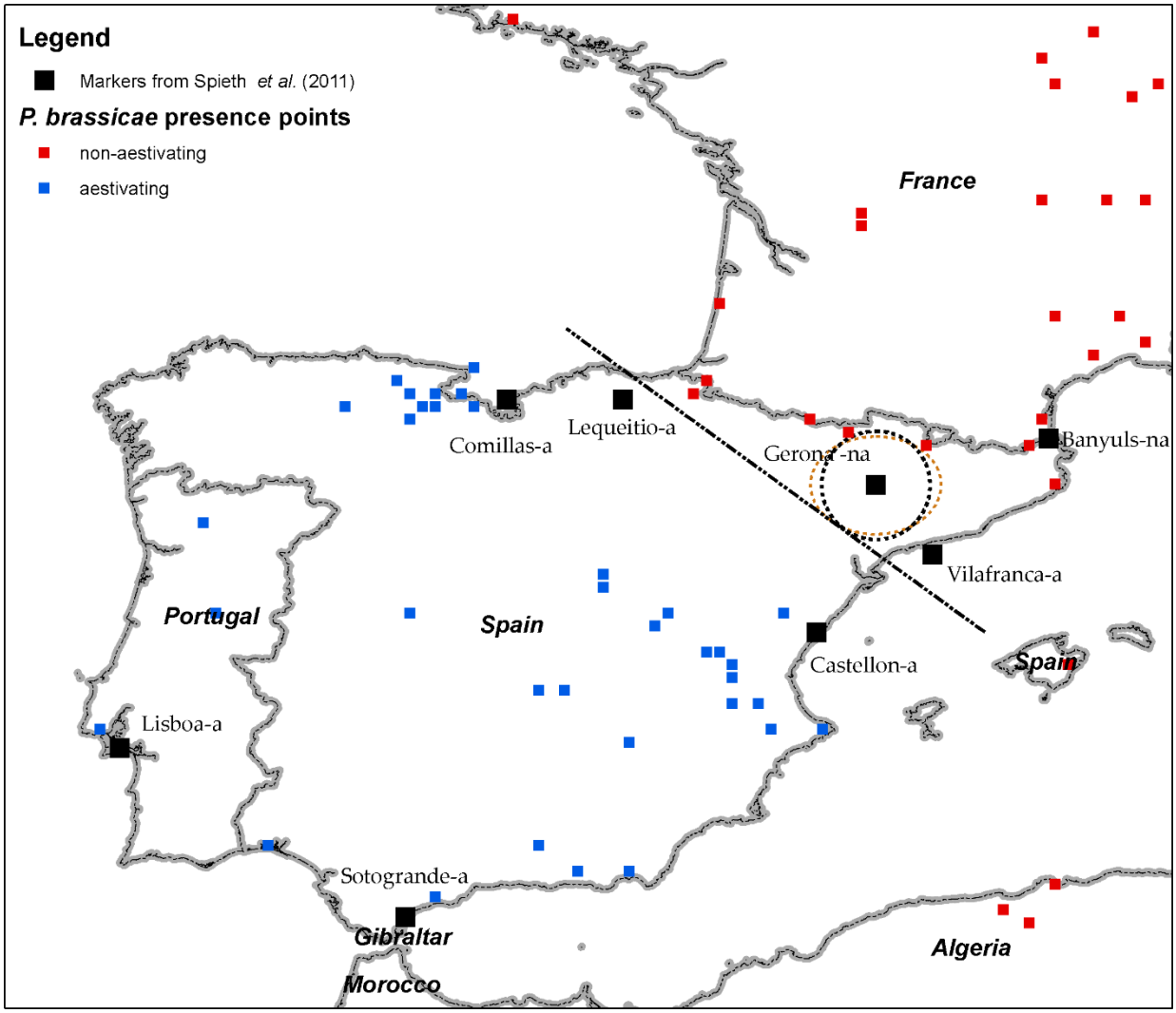


Figure 5.2 Classification of P. brassicae *presence points. The diagonal black line shows the cline where* P. brassicae *populations transition from non-aestivating to aestivating types. The black squares show spatial markers (place names) used to describe the geographical boundary of* P. brassicae *by Spieth et al. (2011). Labels show place names along with an "a" or "na" suffix which means aestivating or non-aestivating respectively. It was reported that* P. brassicae *populations were not aestivating in Gerona even if aestivating populations were found 70 km south of Gerona. Accordingly a circle around Gerona was drawn with 70 km radius to mark a point through which the transition should pass while being north of Lequetio and tangent to the circle at the same time. While this left Vilafranca, where aestivating populations were reported out of the aestivating side, I proceeded with the above line as it satisfies all the other descriptions.*

Directional distribution standard deviation ellipses were used to assess the proximity of the New Zealand invaded locations to the aestivating, non-aestivating and combined presence points. Directional distribution ellipses are usually used to assess central tendency, dispersion and directional trends of spatial features (Lefever, 1926). The derivation of the standard deviational ellipse has been improved by Furfey (1927) to use Cartesian co-

ordinates, and a further improved derivation of the directional ellipse of a spatial data distribution was given by Gong (2002). The naming of the standard deviational curve in geographical space as an "ellipse" has been questioned by both Furfey (1927) and Gong (2002), as other geometrical forms of the curve were obtained depending on the spatial dispersion of the distribution of a given data. However referring to the standard deviation ellipse (SDE) as the standard deviation curve as suggested by Gong (2002) confuses it with the familiar standard deviation curve usually used for the bell shaped normal standard deviation distribution. Therefore, the SDE is referred as an ellipse in this study. Major spatial analysis software including ESRI®'s ArcGIS also still refer to the SDE curve as ellipse.

I adopted this method to assess the proximity of the New Zealand *P. brassicae* locations to both aestivating and non-aestivating presences in the feature space of variables selected according to aestivating and non-aestivating presences. Directional ellipses are used for spatial data, where autocorrelation is assumed to decline as the distance between points increases. The principal component values used to construct the feature space were also based on continuous environmental variables that co-vary in the environmental space fulfilling the assumption for the use of SDEs.

The SDEs for the aestivating and non-aestivating *P. brassicae* classes were derived from the parameters of presence points distribution on the PCA transformed environmental feature space. Three types of feature space were tested, the first two based on variables selected according to aestivating and non-aestivating presence points respectively, the third feature constructed based on variables selected for the unclassed presences (complete presence data).

The directional standard deviation ellipses for the presence points in the aestivating and non-aestivating class as well as for the complete presence dataset (with no classification) was constructed as follows.

Let, X_i and Y_i denote the value of a presence point from a given presence class on the X and Y axes respectively, where X and Y are the first (PC1) and second (PC2) principal components of the feature space constructed out of the PCA transformed environmental variables.

The standard deviations of each presence class according to the X and Y axes are given by

$$S_x = \sqrt{\frac{\sum(Xi-\bar{X})^2}{n}}, \text{ and } S_y = \sqrt{\frac{\sum(Yi-\bar{Y})^2}{n}} \quad \text{----------eq5.6}$$

Where n = number of points in the presence class, and $\bar{X}$ and $\bar{Y}$ are the mean centres of all the points in the presence class on the X and Y axes respectively.

To construct the standard deviational ellipse at the direction of maximum standard deviation, standard deviations S_x and S_y, obtained in Eq. 5.6 are rotated by an angle θ at the mean centres $\bar{X}$ and $\bar{Y}$. The angle θ is determined by selecting the angle that maximizes the resultant standard deviations S_x and S_y. A simplified formula for the determination of the angle θ as well as the derivation of the rotated standard deviations σ_x and σ_y based on the angle θ was given by Mitchell (2005) and these are given in Eq.5.7 – 5.9.

$$\tan\theta = \frac{a+b}{c} \quad \text{----------eq5.7}$$

Let $\bar{x}$ and $\bar{y}$ be the deviations of the points on the transformed x' and y' coordinate from the mean centre.

Then,

$$\text{a} = \sum(\bar{x}_i{}^2 - \bar{y}_i{}^2),\ \text{b} = \sqrt{\left(\sum \bar{x}_i{}^2 - \bar{y}_i{}^2\right)^2 + 4\left(\sum \bar{x}_i\bar{y}_i\right)^2}, \text{ and } \text{c} = 2\sum \bar{x}_i\bar{y}_i \quad \text{-----eq5.8}$$

Using the relationship between Eq. 5.7 and Eq. 5.8 the standard deviations in the rotated axes are given by Eq. 5.9.

$$\sigma_x = \sqrt{2}\sqrt{\frac{\sum(\bar{x}_i\cos\theta-\bar{y}_i\sin\theta)^2}{n}} \text{ and } \sigma_y = \sqrt{2}\sqrt{\frac{\sum(\bar{x}_i\sin\theta-\bar{y}_i\cos\theta)^2}{n}} \quad \text{----------eq5.9}$$

Thus, the centroid of the ellipse is at $\bar{x}$ and $\bar{y}$, and $2\sigma_x$ and $2\sigma_y$ are the long and short axes of the ellipse.

Two directional standard deviational ellipses (1SD and 2SD) were derived for each presence data class Pb_{Aes} (n=35) and Pb_{Naes} (n= 2,206) as well as the unclassified *P. brassicae* dataset (n= 2,241). The SDEs were computed using spatial statistics extension in ArcGIS.

The direction of the ellipse was also used as an additional measure to determine the direction of the different presence data distributions in the feature space. As the standard deviations of the distributions studied in the feature space are the long and short axes of the ellipse, let the values obtained in Eq. 5.9 be given as σ_{short} and σ_{long}. In the event these two values are equal the distribution is circular or uniformly distributed in all directions.

To determine whether the distribution is directional in the feature space, I used the recommendation by Gong (2002) to obtain the circularity index (C_i) of the distribution by using the ratio between the two axes. Smaller values show directionality (oblong ellipses) whereas values closer to one show that the distribution is circular. The same assumptions were extended for features on the variable space as the ones used to implement the SDE in a geographical space. Additionally, the exact direction of the ellipse was depicted on the plots by drawing a straight line through the mean centre of the ellipses at an angle θ determined in Eq. 5.7.

$$\mathrm{Ci} = \frac{\sigma_{short}}{\sigma_{long}} \quad \text{eq5.10}$$

5.2.3 Merging predictions (both case studies)

The model with the highest Kappa was selected for each of the four scenarios (Dvv_I, Dvv_N, Pb_{Aes} and Pb_{NAes}). The predictions from the selected models for Dvv_I and DVV_N were combined to produce a final *D. v. virgifera* global and New Zealand extent potential distribution. Similarly, the selected models for Pb_{Aes} and Pb_{NAes} were combined to produce a final global and New Zealand extent potential distribution for *P. brassicae*.

To facilitate combination of predictions from the best models, the point datasets that have the predicted values from the selected models were converted to raster datasets using cell size 10′ and 30″ for the global and New Zealand extent respectively. All "no data" values were set to 0. The respective rasterized predictions from the two components of the *D .v.*

virgifera and *P. brassicae* were combined using the rule given on Eq. 5.11. The map algebra function in the Spatial Analyst extension of the ArcGIS software was used to combine the component raster for both species.

$$Con\left((Comp1 >= 0.5), Comp1, Con\left((Comp1 < 0.5) \,\&\, (Comp2 >= 0.5), Comp2, \frac{Comp1+Comp2}{2}\right)\right) \text{---------Eq5.11}$$

The Con function in ArcGIS raster calculator facilitates conditional statements. Comp1= the prediction from the component with the large number of presences (Dvv$_I$ and Pb$_{NAes}$). And Comp2 = the prediction from the component with fewer presences (Dvv$_N$ and Pb$_{Aes}$). The simple merging rule above was used to keep predictions from the large presence dataset wherever possible to give precedence to the component with high prevalence, hence the precedence in Eq.5.11 for predictions from Comp1. The combination rule at the same time specifies that areas that are not predicted by the large prevalence class (Comp1) but predicted by the low prevalence class (Comp2) are considered in the final prediction.

Since the individual models corresponding to these component predictions were parameterized and validated separately, a simple accuracy and sensitivity validation was undertaken on the final predictions for both species. Sensitivity is chosen as a performance measure for the combined predictions because maximising sensitivity is more important than other performance measures like specificity or precision when it comes to prediction of potential distribution of invasive species to obtain the maximum possible estimation of where these species might establish.

5.3 Results

5.3.1 Testing for significant variation in presence data- case study 1 D. v. virgifera

A normal distribution fit to the histogram plot of the *D. v. virgifera* presence dataset gave a poor fit to the data (Figure 5.3-A). K-means clustering was performed to identify distinct components in the dataset. The result for the K-means clustering with K=2 on the PCA transformed *D. v. virgifera* presence dataset is shown in Figure 5.3–B. The geographical projection of the clustered presences showed a distinct spatial pattern (Figure 5.3-C).

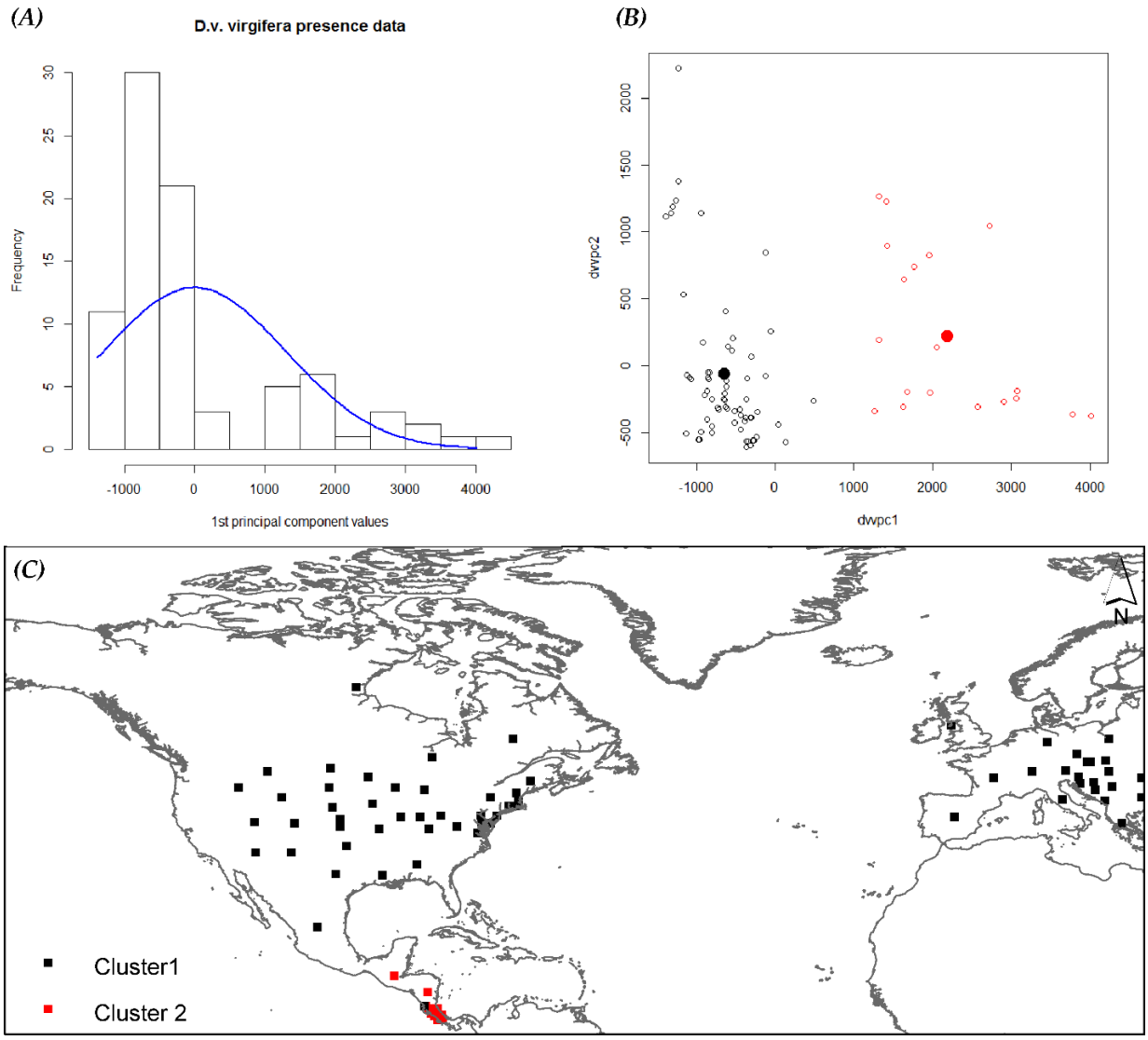


Figure 5.3 D .v. virgifera *presence dataset: (A) histogram plot of the presence dataset with a fitted normal probability density distribution line, (B) the presence dataset after K-means clustering, and (C) the geographic projection of the clustered presence points.*

The likelihood ratio bimodality test (Holzmann & Vollmer, 2008) analysed based on the first principal component values of the presence points in the environmental feature space showed that the mixed normal random distribution parameterized by the two cluster means and standard deviations was bimodal (bimodality test, likelihood ratio = 618.98, $p < 0.00001$, Figure 5.4).

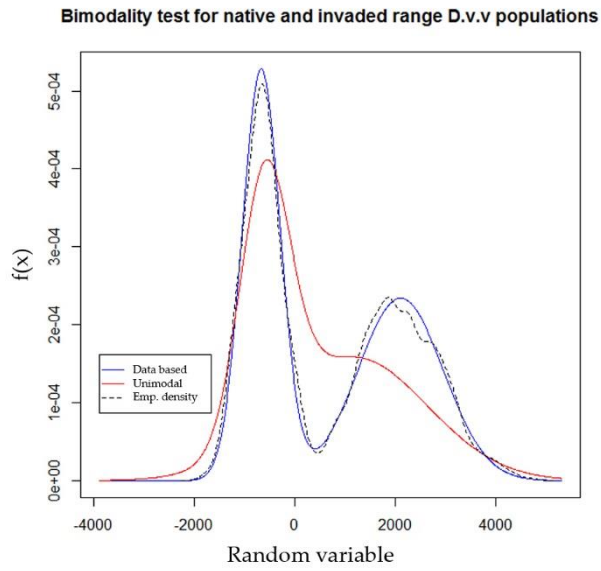


Figure 5.4 the cluster means $\bar{I}$ and $\bar{N}$ and standard deviation S_I and S_N fitted to a mixed random probability density curve (blue line).
The black dotted line shows the empirical density, and the red line shows the unimodal fit.

***5.3.2 Testing for significant variation in presence data- case study 2* P. brassicae**

Because the presence points identified to represent the aestivating population of *P. brassicae* were a small percentage (1.5%) compared with the total number of presences, the method proposed for the *D. v. virgifera* case was not appropriate to assess any variation within the *P. brassicae* presence dataset according to aestivating and non-aestivating presences. A different approach that considers the difference in variables selected when using the two classes of *P. brassicae* presence dataset was employed. This approach was appropriate as it was less density dependent and the values inferred from the presence points and their position in the feature space of the selected variables was important rather than the number of presences in each class.

The relative positions of the aestivating and non-aestivating class presence points with regard to the newly invaded locations in New Zealand were compared in the PCA transformed feature space of four different variable combinations (Figure 5.5).

The first plot (A) where all variables are indiscriminately used did not provide a very good discrimination between the background (the rest of the world) and the presence points. Plot (D) shows the feature space constructed out of variables selected for aestivating presences, here there was a distinct clustering of presence points in the aestivating and non-aestivating classes that is not captured in Plot (B) and Plot (C).

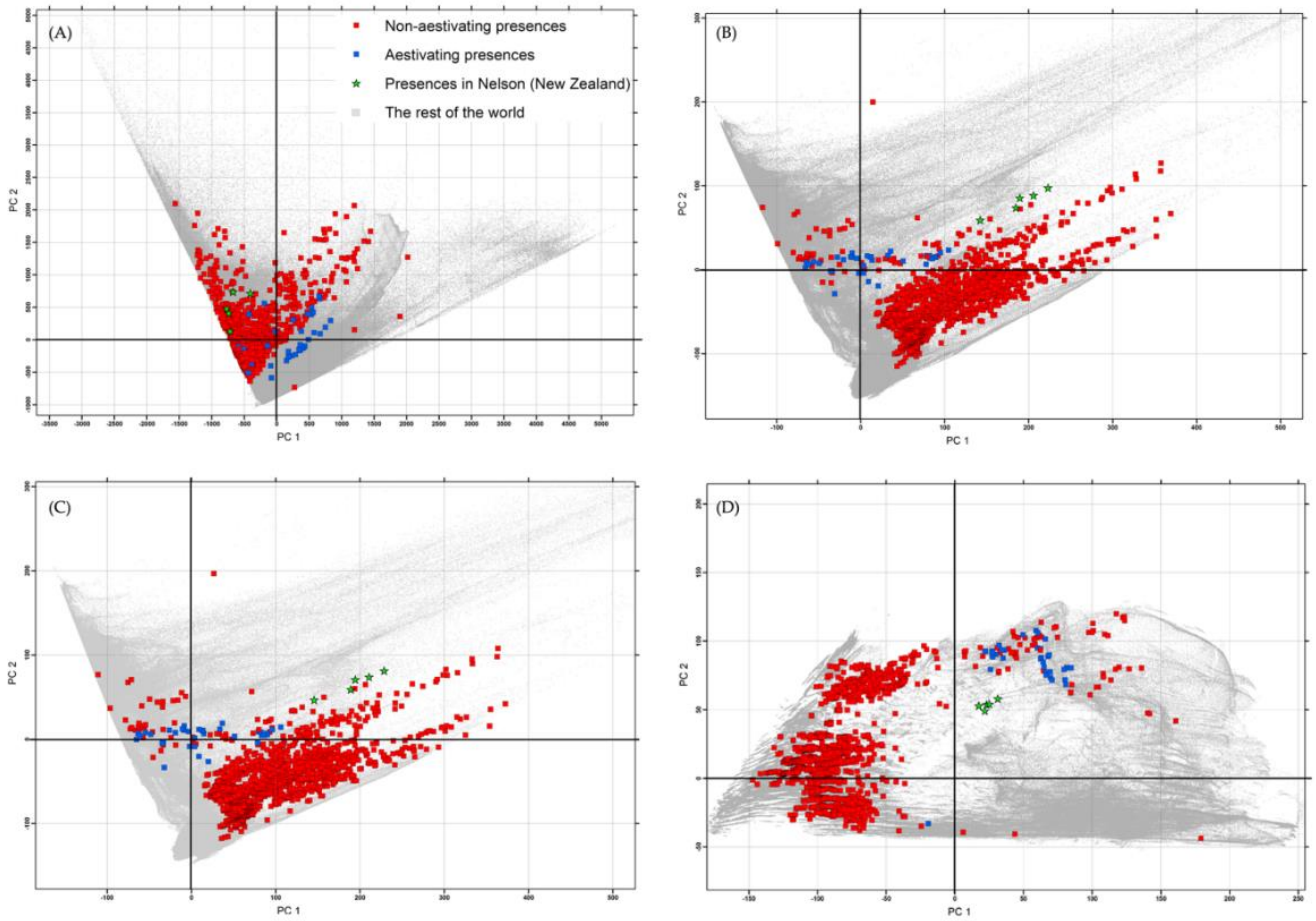


Figure 5.5: Comparison of the relative positions of aestivating and non-aestivating presence points in feature space of four different bioclimatic variables combinations.
(A) P. brassicae *global occurrences in PCA transformed feature space of 39 predictors. (B)* P. brassicae *global occurrences in PCA transformed feature space of 15 variables selected according to the complete* P. brassicae *presence dataset (n=2,241). (C)* P. brassicae *global occurrences in PCA transformed feature space of 11 variables selected according to presences in the non-aestivating class (n=2,206). (D)* P. brassicae *global occurrences in PCA transformed feature space of four variables selected according to presences in the aestivating class (n=35).*

A more systematic analysis was undertaken to check if the effect of the low prevalence aestivating class of presence points, could have been masked when prediction was performed using all presence points. Figure 5.6 shows the one standard deviation (SD) and two standard deviation (2SD) ellipses drawn with the mean centres for the aestivating, non-aestivating and unclassed presences as the respective centre of the directional ellipses.

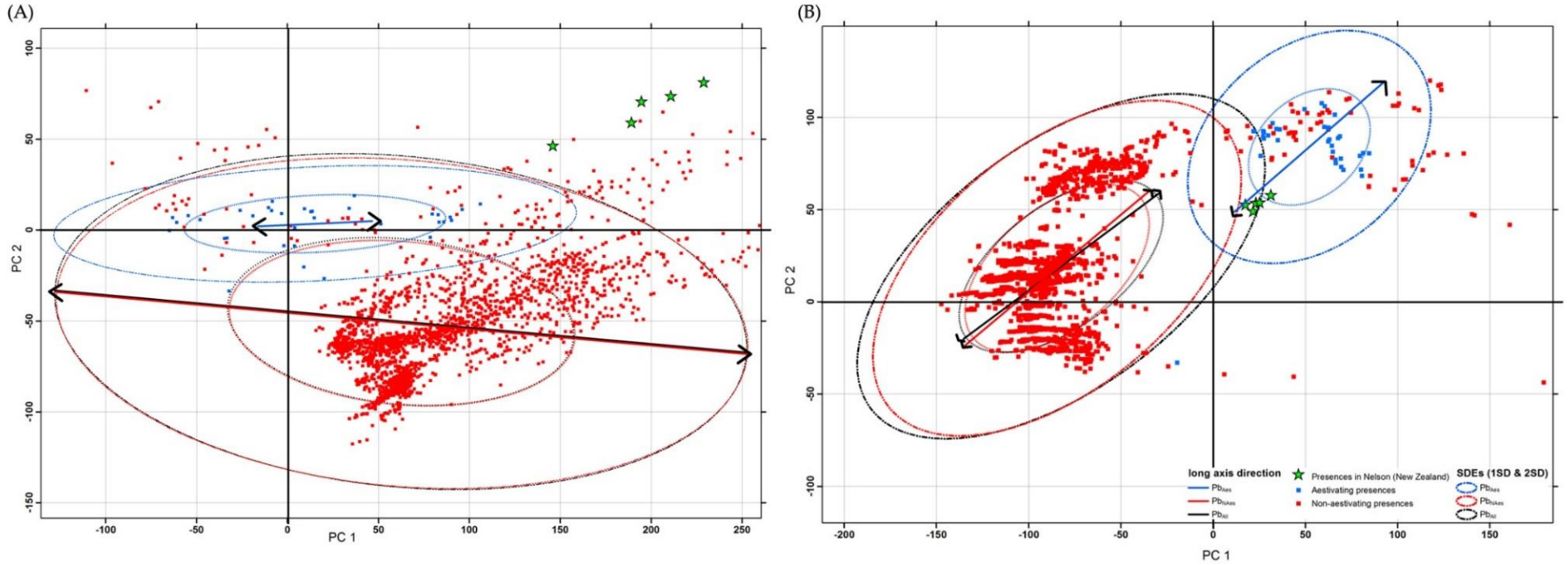


Figure 5.6: Distribution directional 1SD and 2SD standard deviational ellipses (SDE) derived from the centre means of aestivating, non-aestivating and combined presences of P. brassicae *on the feature space of variables selected according to (A) non-aestivating presences (B) aestivating presences. Green stars show* P. brassicae *locations in New Zealand.*
Only the feature spaces constructed out of the variables selected based on the aestivating and non-aestivating class of P. brassicae *presence points were used for the SDE analysis. The feature space constructed out of the variables selected based on the combined presence dataset (Figure 5.6-B) is not considered as it is very similar with the non-aestivating feature space. The long axes of the ellipses indicate the direction of the respective distributions.*

The configuration of the different presence points on the feature space from variables selected based on the non-aestivating presences (Figure 5.6-A) show no proximity between the newly invaded New Zealand locations and *P. brassicae* presence points. The New Zealand locations were outside the 1SD and 2SD ellipses of the aestivating, non-aestivating and combined presence clusters in Figure 5.6-A. The second feature space (Figure 5.6-B) however, shows that the New Zealand locations were partially contained in the 1SD ellipse derived from the mean centre of the aestivating clusters and wholly contained in the 2SD ellipses of all three clusters. The 1SD ellipses of the combined presence points and the non-aestivating points did not include any New Zealand points and was distinctly further from the 1SD ellipse based on the aestivating presences.

Evidently, the feature space according to the aestivating presence points explained the environmental similarity between the invaded New Zealand locations and all *P. brassicae* points better than when all presences or just non-aestivating presences were used to select variables. Moreover, the direction of the distribution of both presence classes was aligned with the newly invaded locations in New Zealand in the second feature space constructed with variables selected for aestivating presence points.

Table 5.2. Circularity index of the directional standard deviational ellipses computed for the three types of presence data classes on two types of environmental variable feature spaces.

Presence data class	Feature space*	mean x	mean y	σ_x (2SD)	σ_y (2SD)	rotation (θ)	Ci
aestivating	1	15.12	3.44	31.39	143.94	87.36	0.22
non-aestivating	1	62.96	-51.19	190.85	89.74	95.12	0.47
all presences	1	62.20	-50.32	191.10	90.88	95.71	0.48
aestivating	2	52.14	83.97	53.57	73.87	49.03	0.73
non-aestivating	2	-84.65	18.21	65.14	118.28	50.03	0.55
all presences	2	-82.46	19.25	66.02	128.87	53.27	0.51

**The Feature spaces one and two are made up of the 1st and 2nd principal components of the PCA transformed data of variables selected based on non-aestivating and aestivating presence points respectively. The shaded rows show ellipses that have higher directional (oblong) distribution, hence the low circularity index (Ci) index. The σ_x and σ_y are given as 2xSD divide the values by two for the standard distance (standard deviation) of the distribution according to x' and y' axes.*

The circularity index for the ellipses of the presence dataset clusters tested is given in Table 5.2. The direction of the different presence distributions on the two feature spaces shown in Figure 5.6 are indicated by the straight line drawn through the mean centre of the respective presence distributions inclined at the angle of rotation of the directional ellipse, this line is also the long axis of the respective ellipses.

5.3.3 Model performance and muli-model comparisons

For *D. v. virgifera*, according to model Kappa scores the models SVM and CART were selected for predictions of the native and invaded ranges respectively (Figure 5.7-A&B). Both of the best models selected for *D. v. virgifera* had acceptable area under the ROC curve scores (AUC > 0.7) (Figure 5.7). The minimum Kappa score from the best models was 0.7, which was within acceptable range to perform prediction.

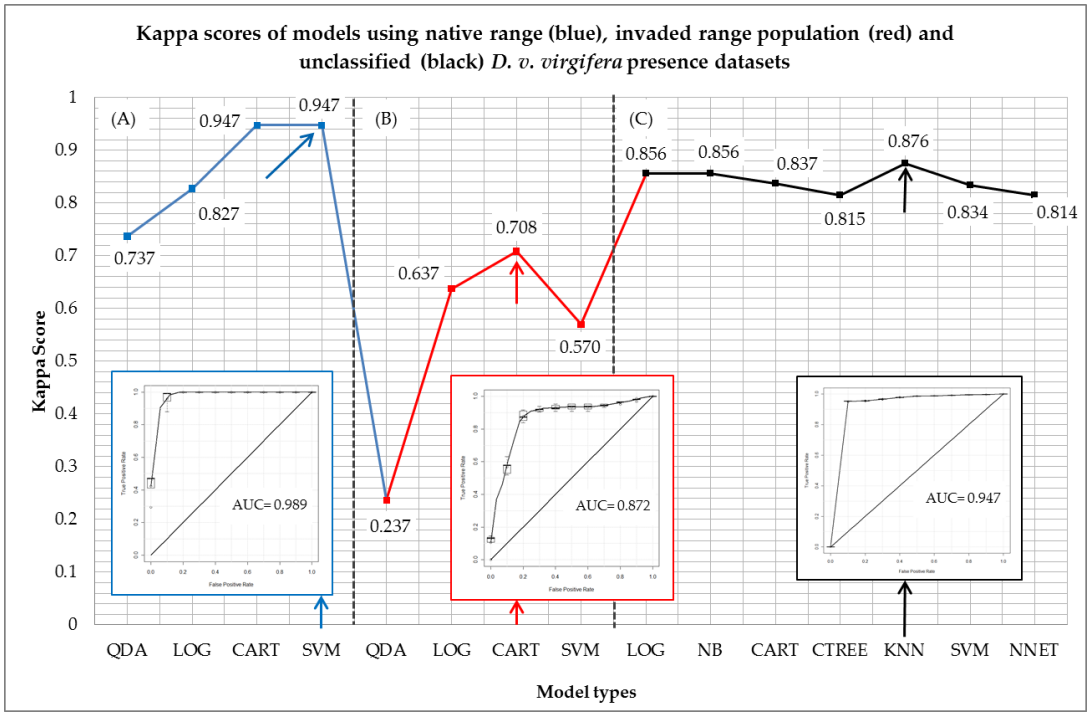


Figure 5.7: Performance of models trained based on presences from the native (A), invaded (B) and unclassed presences (all presences) (C) of D. v. virgifera *populations*
NNET model scores were removed from the sub-plot (A) and (B) to make the graph easily readable as they were almost identical with the SVM predictions in both cases. The model predictions given for (C) are the exact results from the study in Chapter 2 and are shown here for comparison. The arrows show the models with the highest AUC and Kappa scores.

For *P. brassicae*, the Kappa scores of the models based on the unclassified presence points of *P. brassicae* (Figure 5.8-C) were close to the Kappa scores of the non-aestivating presence models (Figure 5.8-B). The scores for the aestivating presence models (Figure 5.8-A) was generally low. Considering only 1.5 % of the presence data was labelled for the aestivating class, the low model performance was expected. However, the best model selected for the aestivating presence models had an acceptable performance (Kappa=0.7, AUC = 0.7) allowing the prediction to be used for the final combined prediction. LOG and SVM were selected as the best models for the aestivating and non-aestivating presence classes respectively.

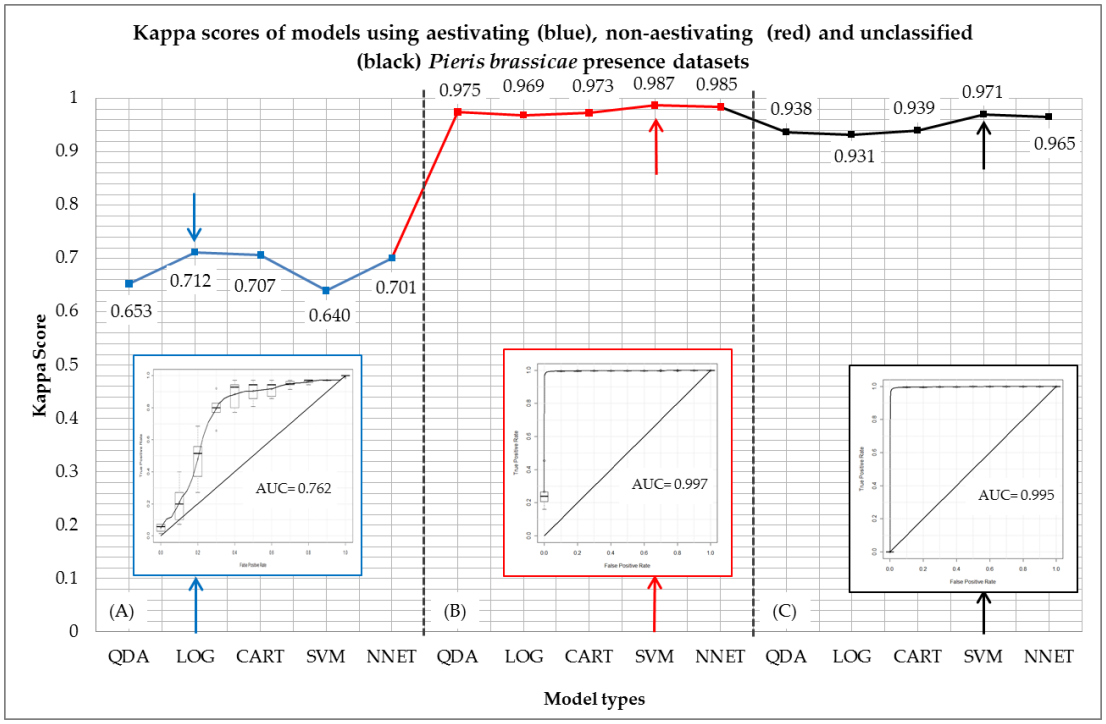


Figure 5.8: Performance of models trained based on presences from the aestivating (A), non-aestivating (B) and unclassified (all presence points) (C) P. brassicae *populations. The arrows show the models with the highest AUC and Kappa scores.*

5.3.4 Combined predictions

The combined prediction for both species was assessed using the test data that was used for the individual component models. Ideally, external validation data would provide better validation for such composite predictions. With external validation one can appropriately assess if distinct areas that were masked due to direct training on mixed component data were identified through separate modelling of these components. Because such data was

unavailable the validation using the same test data was done to check if the combined prediction has a comparable accuracy with the individual predictions.

The overall accuracy (the ratio of correctly predicted presences and absences out of the total number of test points) and sensitivity of the combined prediction as well as the individual components for both species are given in Figure 5.9-A & B.

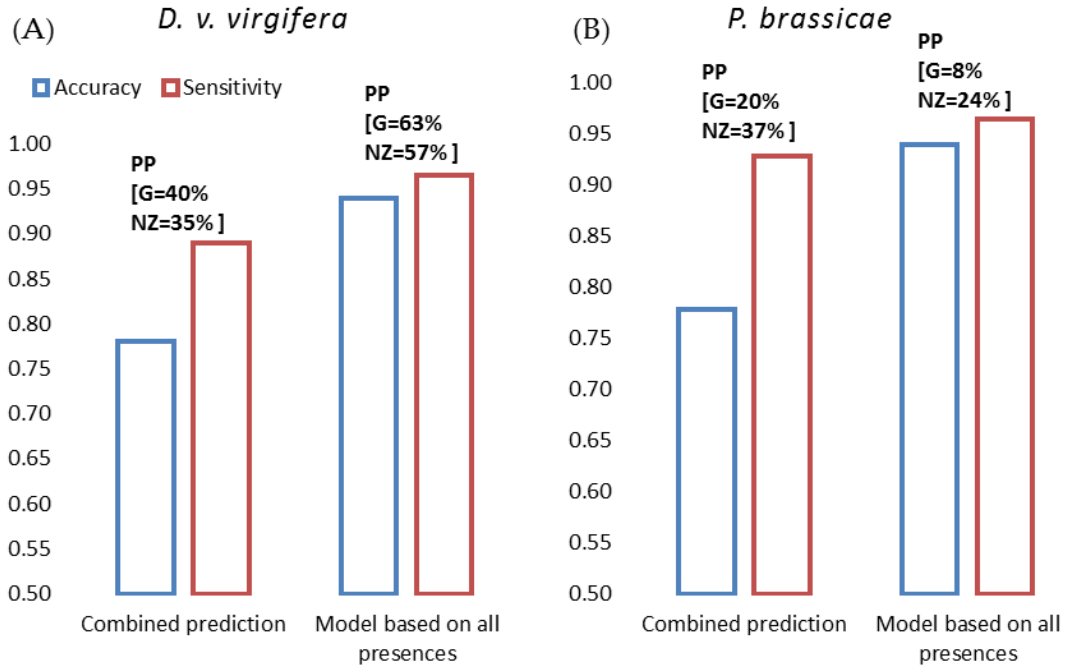


Figure 5.9: Accuracy and Sensitivity scores for the combined prediction for both species using the test data from the component predictions (A) D. v. virgifera *(B)* P. brassicae.
PP in the sensitivity score bars stands for predicted prevalence; the PP % was given both for global (G) and the high resolution New Zealand extent predictions (NZ).

The global *D. v. virgifera* potential distribution prediction obtained when the SDM was trained directly on both native and invaded range presences is shown in 5.10-A. The global potential *D. v. virgifera* distribution prediction, which was the combination of the separate native and invaded range predictions is given in Figure 5.10-B.

The global extent *P. brassicae* potential distribution prediction obtained when the SDM was trained directly on both aestivating and non-aestivating presences is shown in 5.11-A. The New Zealand extent *P. brassicae* potential distribution prediction, which was the

combination of the separate aestivating and non-aestivating population predictions is given in Figure 5.11-B.

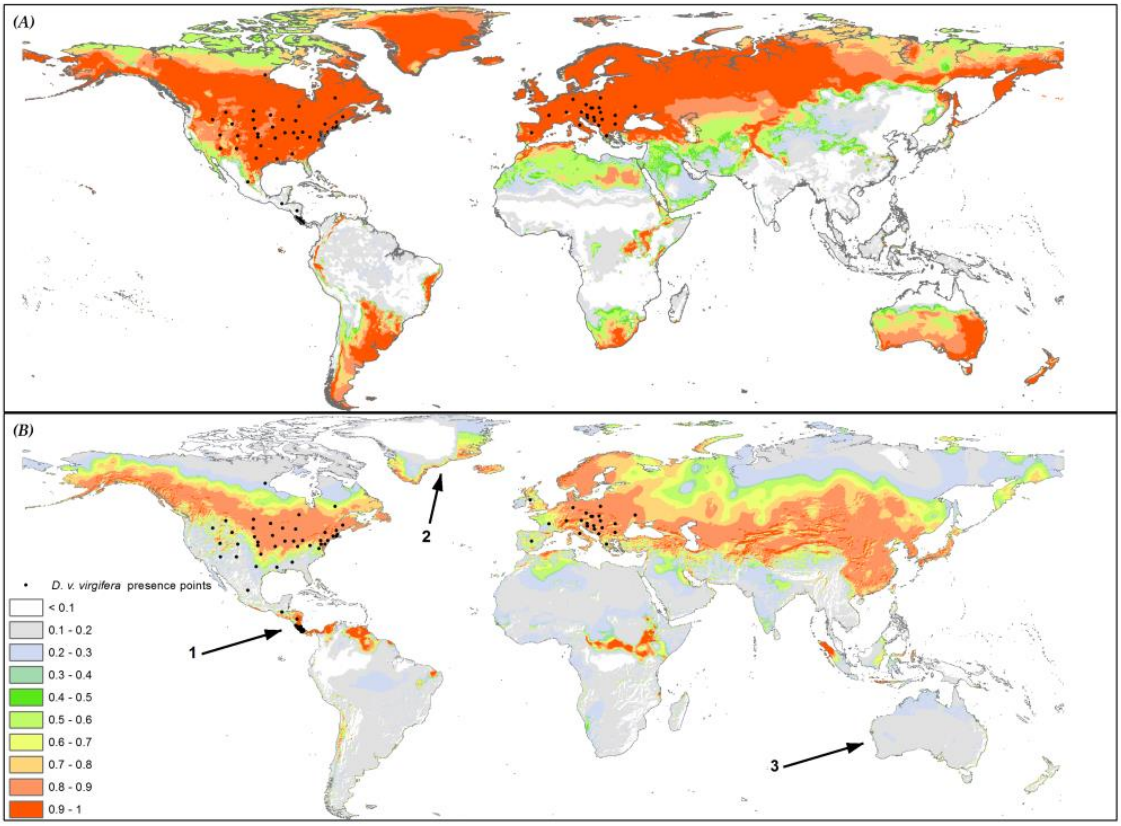


Figure 5.10: Global potential D. v. virgifera *distribution (A) direct prediction (B) combined prediction*

(A) Shows the direct prediction result where the model was trained on presence points from both invaded and native range. The result was taken from the study in chapter 3, K-nearest neighbours (KNN) was the best model out of a seven model framework for the D. v. virgifera potential distribution prediction using the 3-step pseudo-absence method (see Figure 5.7-C for the model performance results). (B) Shows the combined prediction obtained from two separate predictions based on the native and invaded range presences respectively. The numbered arrows show major differences between the two maps, Arrow 1- shows that the combined map predicted the Central American range as well as the invaded range in N. America and Europe, while the first map predicted only N. America and Europe. Arrow 2 shows that the over prediction of suitable areas for D. v. virgifera in the ice covered areas of Greenland in the first map is reduced when predictions are combined. Arrow 3- Even though south eastern Australia was predicted as highly suitable in the three studies in this investigation it was not shown as a high probability for establishment in the combination prediction. Since the species has not reached that part of the world it is difficult to validate the large discrepancy for this region. However, the map of individual predictions (Appendix 5.1) show that the New Zealand prediction was influenced by the presences in the invaded range rather than the native range. Therefore for New Zealand it might be more useful to either use the direct prediction where models are trained on both native and invaded ranges simultaneously, or using presences in the invaded range.

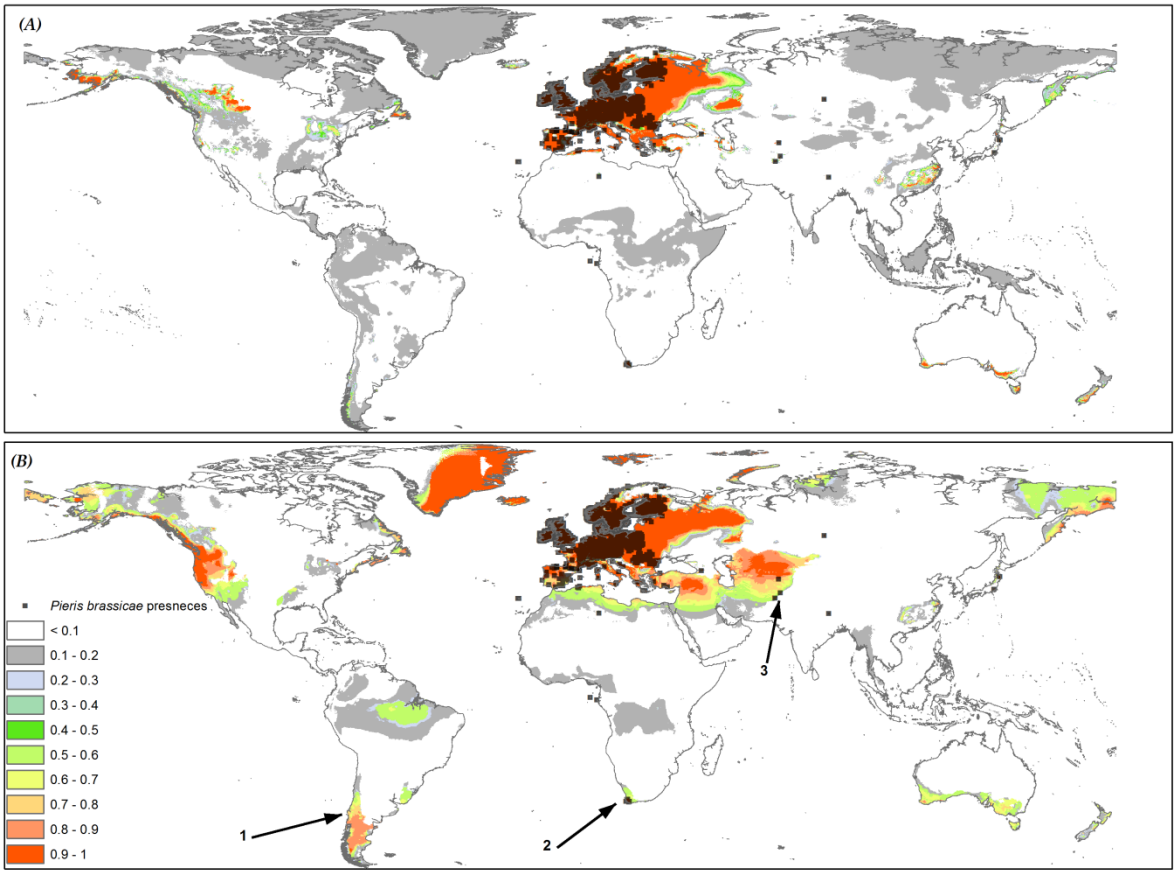


Figure 5.11: Global potential P. brassicae *distribution (A) direct prediction (B) combined prediction. (A) The global direct potential species distribution prediction for* P. brassicae *by the best model (SVM) that was directly trained on both aestivating and non-aestivating presences. Even though the model predicted the European* P. brassicae *range it did not include the newly invaded locations in New Zealand (better shown in Figure 5.12). The predictions for the invaded area in Chile and the Middle East did not also show adequate intensity (predicted presence probability (P) < 0.5). (B) Prediction for* P. brassicae *based on the combined prediction of the aestivating and non-aestivating individual predictions. Note that there were more areas predicted as suitable in the Americas and Africa as well as Australia. The combined model predicted the occurrences in Chile (arrow no. 1), the Middle East as well as the newly invaded locations in New Zealand. Arrow no. 2 shows the invaded location in South Africa. The prediction for South Africa lies in a very limited spot in Cape Town same as the direct prediction above, indicating that the* P. brassicae *was introduced in an environmental suitability island where immediate areas are unsuitable for the species. This could be the possible explanation for the delayed spread of* P. brassicae *up wards into the continent. Arrow no. 3 show high suitability areas in the Middle East that were not predicted in the first map.*

The New Zealand extent *P. brassicae* potential distribution prediction obtained when the SDM was trained directly on both aestivating and non-aestivating presences is shown in 5.12-A. The New Zealand extent *P. brassicae* potential distribution prediction, which was the combination of the separate aestivating and non-aestivating population predictions is given in Figure 5.12-B.

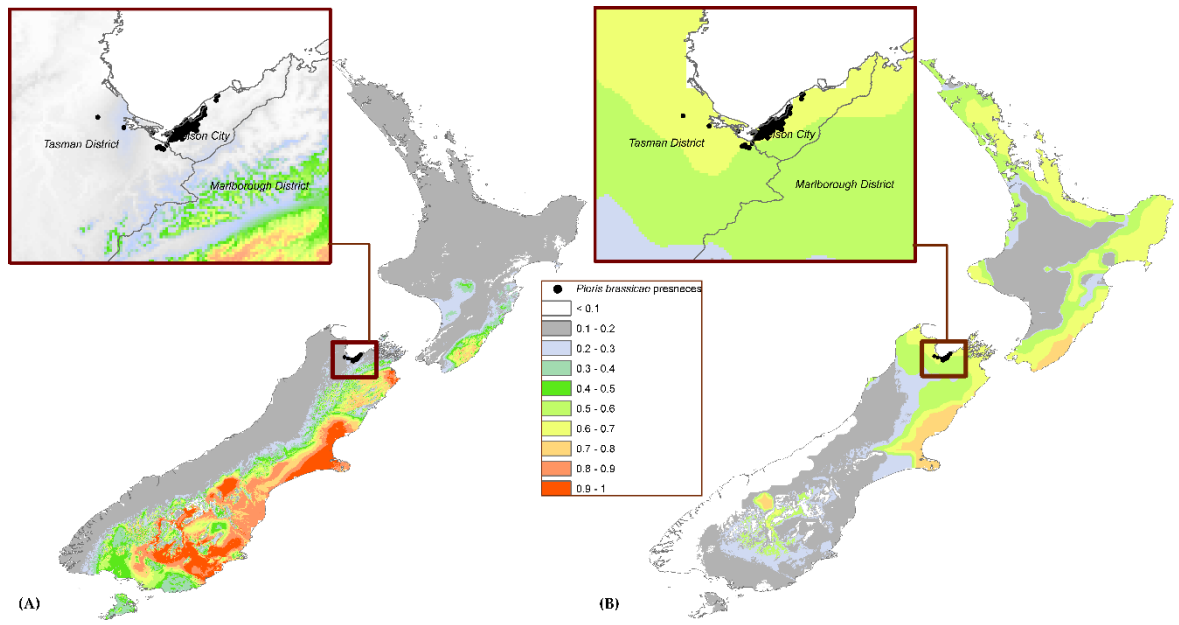


Figure 5.12: Potential P. brassicae *distribution prediction for New Zealand (A) direct prediction (B) combined prediction*

The core distribution of the newly invaded location in New Zealand was not predicted by the model directly trained on both aestivating and non-aestivating presences (Figure 5.12-A). The combined prediction however, has identified all the areas where the core and satellite populations of *P. brassicae* were located (see inset maps of Nelson city and the Tasman district of The South Island, Figure 5.12). The difference in prediction was large for The North Island where more areas were found suitable using the combined predictions (Figure 5.12-B). The predicted presences from the direct and combined predictions were compared with true presences by plotting both on the environmental feature space (Figure 5.13).

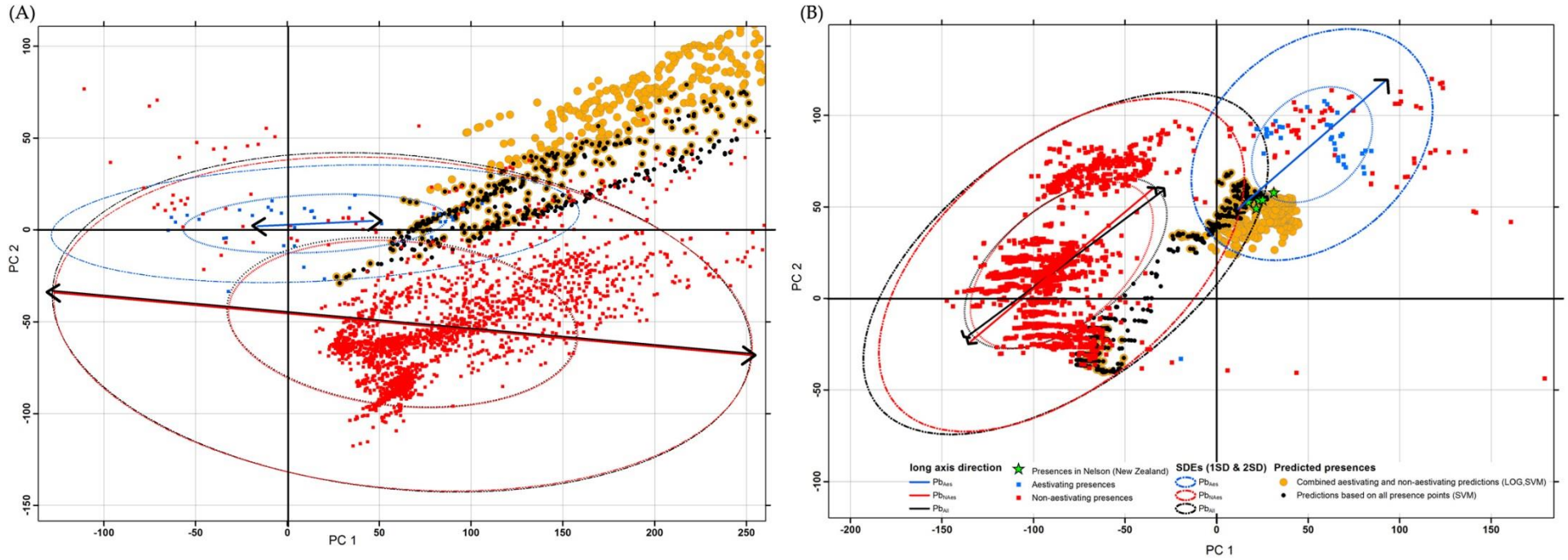


Figure 5.13 Comparison of predicted presences with true presences in the environmental spaces
(A) Combined and direct potential distribution prediction for P. brassicae *in New Zealand plotted in the feature space of variables selected according to non-aestivating presences. (B) The same predictions plotted on the feature space of variables selected based on aestivating presences. The first feature space is also similar with the feature space obtained from variables of all combined presence points thus it is not given here. The fact that the direction of the maximum deviational ellipse is away from the predictions in the first map indicates that the variables used for this feature space were not the most appropriate for the presence dataset. Feature selection models could select in appropriate variables if there is considerable noise in the dataset under study (Steyerberg et al., 1999; MacNally, 2000). The use of a presence dataset with varying components (e.g. aestivating, non-aestivating) can be the reason for the selection of less appropriate variables in the case of the first feature space*

5.4 Discussion

Previous studies showed that correlative SDMs perform best as more presence data becomes available (Urban *et al.*, 2007; Beaumont *et al.*, 2009; Dupin *et al.*, 2011). Additionally, geographical accuracy of the presence points have also been shown to affect SDM predictions (Rodda *et al.*, 2011), hence there is a need for compromise between having a high quantity of presence points (locations) as well as making sure those points are accurately geo-referenced. The other factor that is closely associated with uncertainty in relation to presence datasets is sampling bias. Utilising bias grids for presence only models (Phillips *et al.*, 2009) and limiting background data for pseudo-absence selection in presence-absence models (Lobo *et al.*, 2010) have been suggested to overcome the effect of sampling bias in presence data.

In this study, I have shown that there is another factor that could affect SDM predictions in relation to presence data. The ideal presence dataset has abundant presence points, accurate geographic references and is devoid of sampling bias or it is corrected for presence sampling bias. However, even then, high variation between presence points associated with local environmental adaptation by a specific population of the species could lead to over- or under-prediction of the potential distribution of the species under study.

Modelling mutli-modal data is a well-researched subject and is not a new consideration (Ashman *et al.*, 1994; Ahmed *et al.*, 2008). In case of mutli-modality, mixture models that each explain the distinct components of the data are usually used for modelling (Chen & Li, 2009). Species distribution modelling studies however usually assume species presence points can be explained by unimodal density distributions (Austin, 2007). And this usually could be the case as many species are present in locations which are environmentally similar.

On the other hand, invasive species are successful establishing populations in their new habitats often because of higher environmental adaptation abilities compared with their non-invasive counterparts. As a result the invaded range of an invasive species does not necessarily have to be similar to the native range. Therefore, it is important to investigate the possibility of mutli-modality in presence data especially if a species distribution model fails to predict areas where the species is known to occur either through field data or when their

presence is assumed due to evolutionary reasons. Most machine learning models can handle such mixture component datasets, however if they are not parameterized appropriately the ability of the models to handle such data will not automatically deliver a higher accuracy in predictions of presence data that is multi-modal.

5.4.1 Investigating multi-modality in presence datasets

5.4.1.1 D. v. virgifera

In the *D. v. virgifera* case study, a pre-determined number of clusters (K=2) was used for the cluster analysis to detect whether there was bimodality in the *D. v. virgifera* presence dataset according to the native and invaded range of the species. Such supervised clustering with a user defined number of clusters has less uncertainty if the source of variation in the presence dataset is already known and predictors that accurately describe the variation are chosen.

For *D. v. virgifera* presences there were two distinct peaks according to the parameters from the two clusters (Figure 5.4) but possible additional multi-modality in the data could be further investigated in future studies when more presence points are available. Such investigation is essential as previous genetic studies on *D. v. virgifera* invasion of Europe (Hummel *et al.*, 2008- & references within) reported that *D. v. virgifera* populations have undergone remarkable adaptations in Europe. Therefore, it is important to check if the reported adaptations are environmentally distinct and whether that can create further distinction between the N. American and the Europe presences for SDM modelling. However, a larger number of presence points are needed to avoid false modes in the dataset introduced due to spurious fluctuations caused by the lack of data.

On the other hand unsupervised clustering could give a unique opportunity to investigate unknown variation in presence data (Jain *et al.*, 1999). Since the predictors that best explain the unknown variation cannot be pre-determined, a robust dimension reduction method as well as cluster number optimization might be necessary. It is also important to acknowledge that there are uncertainties with clustering data regardless of the clustering method used especially if there are not enough presence points to characterise any possible distinct components in the dataset.

5.4.1.2 P. brassicae

Some populations of a species could develop certain environmental adaptations that allow them to occupy geographical areas that are distinct from the rest of the population. Such a scenario is one possible way multi-modality can occur in a presence dataset, as presences taken from the uniquely adapted populations will have environmentally different values than the rest of the presences. This type of variation is difficult to detect unless prior information is available. For example, availability of information on the aestivating population of *P. brassicae* in this study made it possible to identify the distinct components in the presence data. Although while testing a different hypothesis, Stockwell and Peterson (2002) also reported that by separately modelling populations of Mexican birds artificially divided through a given latitude and combining the results gave high prediction accuracy than when they used the whole data.

The second case study of the aestivating population of *P. brassicae* in this thesis is a good example of variation in presence data due to biological adaptation. Detecting variation in presence data due to a distinct biological trait is more complicated. The unique response to different environmental conditions might be subtle and could be a result of unique interactions between environmental variables instead of direct responses to selected variables. Care should also be taken not to mistakenly analyse biological trait adaptations due to biotic factors as a result of abiotic factors.

The example given in this study is a borderline case where the reported aestivation in a specific population of *P. brassicae* has been suggested to be caused by either the need to avoid parasitic attack or to synchronize with populations that go into a lengthy winter diapause in the colder northern Europe range of the species (Spieth *et al.*, 2011). However, it was suggested in the same study that the trait could also be attributed to the warmer and longer photoperiod environmental conditions in Southern Spain that facilitates a high number of generations per year for the species. This means that the time spent in aestivation to avoid parasitism is less costly because sufficient generations are produced even with the time lost for aestivation in this warmer environment, compared with the colder ranges of the species. In other words, a possible abiotic factor for the development of the trait.

5.4.2 The effect of variable selection

It is important to understand the impact of variable selection on the interpretation of the underlying environmental space shared by presence points that are assumed to represent suitable environmental conditions for a species. The potential geographic distribution of any species is a function of this common environmental space at least in terms of the assumptions of correlative modelling (Elith & Leathwick, 2009; Austin & Van Niel, 2011).

Therefore, working with the most appropriate variables for a given species is as important as identifying any cause of variation within presence data. For example, according to the directional SDE analysis (Figure 5.6) the aestivating presences of *P. brassicae* were significantly further from the majority of the remaining presence data when plotted in a feature space constructed out of variables selected according to the aestivating presence locations, however this discrimination was not visible when the various presence data classes were plotted either in a feature space constructed of variables selected based on the unclassified presence points or non-aestivating presence points. In fact, the directions of the maximum standard deviations (indicated by the long axes of the SDEs) of the respective *P. brassicae* classes show that the distribution direction of the aestivating presences is different from the non-aestivating presences when projected on environmental space derived from variables selected according to all presences or non-aestivating presences. This result signifies that variables selected based on the non-aestivating or the combined presence data would not have characterised the information within the aestivating presence locations. That result implies an effective masking of the contribution from the aestivating presences during direct modelling predictions.

On the contrary, the SDEs in the feature space based on variables selected according to aestivating presences, show that: 1) the maximum standard deviational direction is similar for all presence points, showing that any of the environmental conditions between 1SD and 2SD, SDEs are likely to be occupied and the fact that the species is not in equilibrium, 2) that the newly invaded New Zealand locations were perfectly aligned within the standard deviational direction of distribution of both presence data classes as well as the overall presence data distribution, where it is contained in the 2SD SDE of both aestivating and non-aestivating presence classes (Figure 5.13-B).

SDE ellipses were shown to be predictive when assessing appropriate environmental variables for *P. brassicae*. All the predicted presences from the direct prediction for New Zealand (no presence data classes) fell into the 2SD ellipse of all presence data classes. And all the predicted presences from the combination model fell in either of the ellipses of the aestivating, non-aestivating or unclassed presences (Figure 5.13-B). That shows that some areas in New Zealand were only predicted by the combination prediction and would have been missed if the direct model that is trained in all presences was used. Testing the use of the SDE with more species presence data of known within variation is recommend before the method can be accepted as a general variable selection optimization tool.

Another important point to be made on the use of SDEs for such analyses is that the variables might have complex and non-linear interactions, and since SDEs are based on a linear statistics they may not be appropriate for all cases. However, an effective method could be to use a non-linear PCA or other non-linear dimension reduction methods that constructs the feature space using different non-linear functions instead of the direct linear relationship assumed in the PCA. Then one could undertake the SDE on the resulting feature space.

Provided that a statistical method rather than expert knowledge is used to perform variable selection from the provided predictor data, the type and number of variables selected entirely depends on the training dataset and the algorithm used for variable selection. That also means that the composition of the training dataset (presence and absence points) directly affects the variables chosen. If there are mixed components within the presence data that are likely to be explained by significantly different environmental variables as shown in this study, it is important to subscribe the appropriate variables for each component by separately considering the various presence data groups that correspond to a distinct population of the species.

Morency *et al.* (2010) provided an interesting methodology on joint feature selections for multi-modal data in their study that was designed to provide better human-virtual interaction systems. Such a process could be applied for joint variable selection in multi-modal presence datasets that compare the variables selected for the individual component presence data classes with the unclassified presence data to build a variable set that can

jointly explain all components. Such variable selection could be an alternative to modelling the components separately and combining predictions. But more studies need to be conducted to verify if the jointly selected variables effectively predict areas that would have been predicted if the components were individually modelled.

5.4.3 Combined predictions

A simple rule was set to combine the different component predictions based on the distinct presence data components identified from *D. v. virgifera* and *P. brassicae* presence datasets. Proceeding with predicting the potential distribution of the species according to the different components within the presence data separately, and combining the results later is the simplest and probably the most straight forward method for dealing with mixed component presence data.

However, the rules used to combine the component predictions could be a source of considerable uncertainty as there was no unified performance measure to apply to the combined predictions. For example, the rule to combine the component predictions in this study was set in a such a way that the majority presence data component is given precedence when it comes to assigning values to the final combined prediction and whenever the major component failed to predict an area the alternative component was used to assign prediction values. This rule maximizes sensitivity of the combined predictions which means more environmental variation will be accounted for, compared to individual component predictions. It unfortunately also leads to a low specificity which introduces significant commission error in the combined prediction compared to the individual component predictions.

The comparison between the direct prediction and the combined prediction based on model sensitivity and overall accuracy showed that the direct prediction had better scores (Figure 5.9). Which means for the global extent the direct prediction performed well. However, The combined *P. brassicae* potential distribution prediction correctly identified the invaded regions in New Zealand and Chile (where *P. brassicae* is confirmed to be established, but for which there were no presence points included in the training and test datasets).

These externally predicted presence locations compelled me to investigate the combined prediction further. One reason for the low model performance of the combined models could be the test data used in the assessment. The presence and pseudo-absence points used to test the individual component models were also used for the combination model. While the presence test data does not pose any problem, the pseudo-absences however can impose a very conservative measure for the combined prediction. This is because it is highly likely that pseudo-absences generated for the aestivating population could encompass non-aestivating presences and vice versa. Therefore, a subset of the pseudo-absences generated for the individual components might include presence points in the combined prediction leading to lower overall accuracy. Such drawback can be avoided by setting aside a percentage of presence points from all identified components as a test dataset for the combined prediction, which means these data points should not be used either as a training or test data for the individual component predictions. In this study, it was not possible to set such data aside as one of the components had very few points.

Mixed component modelling is a new practice in SDM analysis, therefore an in depth study and analysis is required to develop sound mixed model evaluation methods to confidently compare direct and combined predictions and decide the better choice case by case, as the choice is likely going to depend on the species data and study extent. In case of this study it is clear that the combined prediction gave better information regarding suitability of New Zealand to *Pieris brassicae*. This is further shown in the SDE analysis (Figure 5.6 A) where the invaded area in New Zealand would not have been predicted in the direct prediction because those locations were closer to the aestivating *P. brassicae* population presences which were left as outliers in the direct prediction.

The use of mechanistic models that depend on independent physiological limits as a test system to validate correlative SDM predictions has been suggested in a number of studies (Kearney *et al.*, 2008; Monahan, 2009; Aragón *et al.*, 2010; Buckley *et al.*, 2010). Clearly, however the physiological limits need to be known and established by experimentation, which is not the case for many invasive insect species.

When physiological limits of a species are known, it may be especially important to validate such combined predictions using physiological environmental thresholds of the target

species, as it is often the areas where the species is not currently established that are likely to contribute to most of the uncertainty in correlative SDM results and particularly in combination prediction results such as demonstrated in this study. For instance, when the combined global potential distribution of *P. brassicae* in this study was compared to the prediction obtained by training models directly on both aestivating and non-aestivating population presences (Figure 5.11), the core distribution of *P. brassicae* in Europe is predicted similarly in both cases. For areas away from the *P. brassicae* native range the combined prediction identified more suitable areas. For example, in the American and African continents, and more locations in Australia and New Zealand (Figure 5.11-B).

There was no information to validate some areas predicted in the combined prediction, for example in Africa and North America. A number of questions need to be answered to validate these new predictions that were not identified with the direct model prediction. Such as, are these new predictions a result of high commission error? Or a result of unmasking the effect of the aestivating presences which were not considered in the direct predictions? Or can it be a combination of both factors suggested above? Such questions can only be answered if the species is introduced into these areas, or mechanistic models based on physiological thresholds are used to independently verify the discrepancy in predictions. Meanwhile, the SDE analysis of the *P. brassicae* presence components in the various feature spaces as well as the improved prediction of the locations where this species has invaded in New Zealand, with the combined model, suggests that within-presence data variation can affect over all potential species distribution predictions.

5.5 Conclusion

It is often difficult to decide if species presence data contains data representing sub-populations adapted to different environmental conditions. Often the presence data that is generally available for most invasive species is limited. Even if multi-modality is observed during data exploration, if the presence data is composed of too few presence points, the apparent multi-modality could be a result of spurious local maxima that are caused by the lack of data, rather than explaining the true nature of the species density curve. Even more confusing, even if a dataset has abundant points, distinct variation among different populations of a species could still be masked if the appropriate variables that explain that

variation are not used to investigate the data. Statistically, heterogeneous populations with a mixture of two distinct components might not necessarily become bimodal (Holzmann & Vollmer, 2008), whereas two components of a homogenous population could have more than one mode due to lack of data.

Therefore, it is important to develop a credible background about the species biology, geography and other environmental associations before assuming bimodality in presence datasets. Otherwise, extra steps taken to pre-process distinct presence data components could cause data dredging at worst or simply complicate a species distribution modelling process unnecessarily. Moreover, such split predictions followed by combination of component predictions is likely to over-predict potential species distributions due to increased commission errors, which might be overly maximized if a unimodal distribution is unnecessarily assumed to be bimodal.

Despite all that, species could adapt to a new habitat either though specialized biological traits like diapause (Spieth, 2002) or by forming new symbiotic, host or predator-prey relationships with other species (Altieri *et al.*, 2010), or simply acquiring higher tolerance to new environmental conditions outside their native range (Lee, 2002). Therefore it is important to carefully investigate response curves of the presence sample dataset of any species before performing species distribution modelling (Sangermano & Eastman, 2007). Such data investigation helps reduce the possibility that certain populations of the species represented by some of the presence locations have evolved a micro-niche that might not apply to all members of the species. This is especially true, if areas where the species is known to occur is not appropriately predicted through the usual direct modelling methods or if there is biological/ biogeographical evidence that shows that the species have populations that occupy environmentally distinct areas from the rest of the population.

If a heterogeneous population in terms of adaptation to different environmental conditions is represented in species presence data, the contribution of one or more of these distinct populations towards the total potential distribution of the species could be masked due to combined processing of all the presence locations. In such cases, it is advisable to model the distinct components of the presence dataset separately and combine the predictions using a set of rules that maximize a selected performance measure depending on the objective of the

study. Additionally, more complex rules can be set to consider multiple performance measures. For example, predictions can be weighted according to the individual component sensitivity, specificity, Kappa, AUC or other model performance scores.

It is also possible to use complex models that consider multi-modality in a presence dataset. Modular associative neural networks have been reported to handle multi-modal datasets satisfactorily (Crepet *et al.*, 2000). However since these have not been covered in this study I can only recommend more research of this aspect, as such models could avoid the need for individual component predictions and facilitate a direct combined predictions for multi-modal presence data as is done in most species distribution modelling.

Chapter 6

6. Hybrid species distribution modelling

6.1 Introduction

A number of improvements to species distribution models including the ones proposed in the preceding chapters have been suggested to increase their prediction accuracy (Stockwell & Peterson, 2002; Elith *et al.*, 2006; Chefaoui & Lobo, 2008; Wisz & Guisan, 2009; Kampichler *et al.*, 2010; Warton & Shepherd, 2010; Lorena *et al.*, 2011; Barbet-Massin *et al.*, 2012; Senay *et al.*, 2013). The continued effort to improve the methods used for modelling species distributions (Araújo & Guisan, 2006; Venette *et al.*, 2010), is essential as these predictions are often used to inform researchers and decision makers where species are likely to establish a viable population and where they may not (Hannah *et al.*, 2002; Guisan & Thuiller, 2005).

However, there are thresholds beyond which results from SDMs cannot be reliable, especially when used to predict species distribution under future climate scenarios (Sinclair *et al.*, 2010). Correlative models assume that environmental values observed at the sites of current presence of a given species can be used as proxy measures of physiological suitability of that area for that species (Guisan & Zimmermann, 2000; Austin, 2007). However, when these models are used to project the observed environmental data onto new environmental ranges or in a future climate scenario there is no evidence the above assumption may hold. Physiological response of species to a set of environmental conditions in their current environment might not remain constant over extreme changes in space or time (Monahan, 2009; Araújo & Peterson, 2012). Therefore, when it comes to projecting species distribution onto new environmental ranges, especially under future climate

scenarios, it is necessary to further investigate physiological limits of the species in order not to extrapolate (Helmuth *et al.*, 2005; Elith *et al.*, 2010; Diamond *et al.*, 2012).

Mechanistic models are offered as an alternative to correlative models for more precise prediction of species distributions without the need for occurrence data. There is much evidence in the literature that shows mechanistic models predict new environmental ranges into which species might expand better than bioclimatic envelope models (Helmuth *et al.*, 2005; Kearney, 2006; Elith *et al.*, 2010). These models range from sophisticated process based models that may account for nutrient intake, energy balance, food web interaction (Buckley *et al.*, 2010), to models that only require the physiological or developmental limits of the species measured in terms of environmental conditions that directly affect survival (Monahan & Tingley, 2012). Some of the direct environmental variables used in the latter models include temperature thresholds, relative humidity level, soil moisture and photoperiod.

Mechanistic models, however also have some shortcomings. First, a number of species specific biotic parameters are required to run these models and this information might not be readily available. Second, environmental thresholds used in model construction might not be the exact limiting factor in the field as interaction between various other proxy environmental, geographical or climatic factors could modify the target environmental variable. This is because values used to calibrate mechanistic models are usually direct measurements often taken in a laboratory under a controlled experiment (Buckley *et al.*, 2010).

Potentially, these two modelling techniques could complement each other. For example, the limitations of the correlative model to identify areas with environmental conditions which are already in the existing physiological tolerance of the species, but are not realized in the current landscape could be covered by predictions from mechanistic models (Morin & Lechowicz, 2008). On the other hand, the limitation of a mechanistic models to appropriately calibrate environmental variables to correspond to physiological limits of a species can be complemented by correlative models that can utilize numerous variables, which allows them to provide good explanatory power to understand which environmental variables limit the species' distribution.

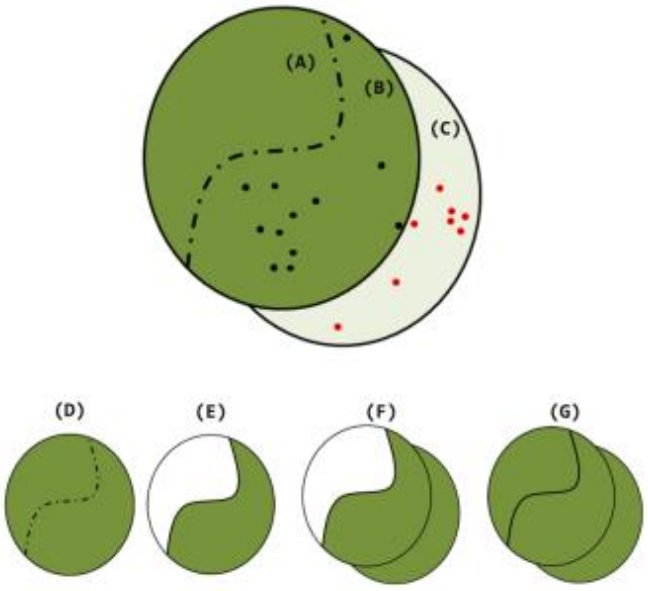


Figure 6.1: Potential species distribution prediction according to correlative, mechanistic and hybrid species distribution modelling methods.
(A) Represents the portion of the species range for which there is no occurrence data; (B) Represents the portion of the species range where occurrence data is available; (C) Represents a new environmental range into which the species is expanding. Black dots show occurrences in known environmental ranges and red dots show occurrences in newly invaded ranges. (D) prediction by mechanistic models; (E) prediction by correlative models with occurrence data only in the existing known environmental range of the species (F) prediction by correlative models with complete occurrence information; (G) Hybrid model predictions shown either as (D+E) or (D+F) depending on the type of occurrence data available. This diagram shows the higher probability of covering more of the species range in hybrid predictions than when mechanistic or correlative models are used separately.

This suggests that even more accurate species distribution predictions are possible by combining these modelling methods instead of their individual use. This idea of using a hybrid correlative and mechanistic species distribution modelling has been proposed by a number of previous reviews and studies (Kearney & Porter, 2009; Elith *et al.*, 2010; Nabout *et al.*, 2012).

6.1.1 Case study: Prediction of the potential distribution of **P. brassicae** *using a hybrid model*

Further validation of the potential distribution prediction made for *P. brassicae* in Chapter 5 was necessary to evaluate if the suitable areas identified in New Zealand were correctly predicted. Ensuring accuracy of the predictions was needed because the suitability map was to be used for an eradication simulation study reported in the next chapter (Chapter 7). *Pieris brassicae* has recently invaded New Zealand (Kean & Phillips, 2013b) where an eradication

process has already been implemented in the area where it established (Kean & Phillips, 2013d).

A number of interesting results about the distribution of *P. brassicae* were obtained from the combined component prediction in Chapter 5. For example, the suitable areas in the north of New Zealand, the south of Australia and northern areas of the South American continent were only predicted with the modelling method that accounted for variation within presence data. In this chapter a hybrid model was used to investigate if these areas were over-predicted or if they were correctly identified suitable habitats.

Most hybrid model studies are based on cases where there is sufficient biotic data to develop mechanistic models, and the results from these mechanistic models can be used in the process of developing a correlative model (Kearney & Porter, 2009; Elith *et al.*, 2010; Nabout *et al.*, 2012). However, in cases where there is no sufficient biological data, it is important to design a mechanistic approach that requires as few biotic parameters as possible. Because it requires minimum parameterization a generalized mechanistic niche modelling method proposed by Monahan (2009) is adapted to construct the hybrid model used in this study.

6.2 Methods

The eighth objective of this thesis was to evaluate the benefits of a hybrid prediction from a correlative model and a simplified mechanistic model as a suitable framework to facilitate improved correlative species distribution predictions. Accordingly, the methods developed to produce a global and New Zealand extent hybrid prediction of potentially suitable areas for *P. brassicae* are given below.

6.2.1 Simple Generalized mechanistic niche model framework

6.2.1.1 Physiological data

Thermal thresholds that directly affect the physiology of *Pieris brassicae* were obtained from the literature (Table 6.1). Lethal and optimal temperature thresholds corresponding to survival and reproduction of *Pieris brassicae* respectively were used to identify the fundamental thermal niche and other physiologically important thermal ranges of the species both globally and for the New Zealand extent.

Table 6.1: Lethal and optimal thermal thresholds of *Pieris brassicae* used in the study

No.	Parameters	Values	Source
1.	Upper lethal temperature (ULt)	40°C	(Wigglesworth, 1945); (Feltwell, 1982)
2.	Lower lethal temperature (LLt)	-26.4°C	(Sømme, 1967); (Hansen & Merivee, 1971) as referred by Turnock and Fields (2005);
3.	Realized thermal range (RTR)*	-20°C to 28 °C	(Klein, 1932) as referred by Feltwell (1982);(David & Gardiner, 1962)
4.	Highly suitable thermal range (HSTR)ǂ	10°C to 26°C	Koehler and Geisenhoffer (2012)

**RTR -this thermal range is based on the winter (January) and summer (July) isotherm of* P. brassicae *reported by Feltwell (1982) ǂHSTR-this thermal range gives the optimum laboratory thermal range in which* P. brassicae *reproduction is maximized and diapause period is minimized.*

The lower lethal temperature used in this study was -26.4 °C as recorded by Sømme (1967) for a diapausing pupae. However, it is important to note that, even the non-diapausing pupae was reported to have a super cooling point of -21.4 °C, that indicates the general high tolerance of *P. brassicae* to low temperatures (Pullin & Bale, 1989; Pullin *et al.*, 1991).

The realized thermal range (RTR) described in Table 6.1 is the temperature range in which *Pieris brassicae* can persist at a location as these isotherms have been consistently reported to describe the presence of *P. brassicae* in the northern hemisphere (Klein, 1932; Feltwell, 1982). The temperature range given as highly suitable (HSTR) is the range within which *Pieris brassicae* was observed to have the highest number of generations, low egg mortality and minimum days for completion of the pupal stage (David & Gardiner, 1962; Koehler & Geisenhoffer, 2012).

6.2.1.2 Climate data

A 50 year average global maximum and minimum temperature data at 10″ resolution was accessed and downloaded from the WORLDCLIM website (Hijmans *et al.*, 2005b). The maximum and minimum temperatures of the coldest and warmest months were calculated using Cell statistics function of the ArcGIS (ESRI, 2010) spatial analyst toolbox. Similar analysis was done for a higher resolution 30′ New Zealand extent WORLDCLIM dataset. The four derived variables used in the generalized mechanistic niche model are given in Table 6.2.

Table 6.2: Variables used in the generalized mechanistic niche model

No.	Variables and parameters	Derived from	Source dataset
1.	Minimum temperature of the coldest month	Minimum monthly temperature	WORLDCLIM*
2.	Maximum temperature of the coldest month	Maximum monthly temperature	WORLDCLIM
3.	Minimum temperature of the warmest month	Minimum monthly temperature	WORLDCLIM
4.	Maximum temperature of the warmest month	Maximum monthly temperature	WORLDCLIM

* http://www.worldclim.org/ ; (Hijmans *et al.*, 2005b)

Determining a generalized lethal temperature for a species that has a varied threshold for each of its different life stages is difficult. This complication can be easily handled by using independent thresholds in process based ecophysiological (Kearney *et al.*, 2008), phenological (Yan *et al.*, 2000) or food web models (Pimm & Rice, 1987). However, for generalized mechanistic niche models such as employed in this study it is important to carefully incorporate certain assumptions to use a single value for the species. The following two sections discuss aspects of the biogeography of *Pieris brassicae* that are considered in order to use the extreme thermal thresholds reported for the species.

6.2.1.3. The realized historical and current geographical distribution of Pieris brassicae

Pieris brassicae has been reported to occur in extremely harsh temperature conditions (Sømme, 1967; Feltwell, 1982). Early occurrence records of the species show its presence in temperature conditions as low as -22.5 °C in its pupal life stage (Bonnemaison, 1965). In contrast, *Pieris brassicae* has also been recorded in warm climate countries like Syria, Lebanon, Israel and Morocco (Feltwell, 1982). The species is known to have a typical Palaearctic distribution prior to its introduction to Chile (Kellner & Shapiro, 1983), South Africa (Gardiner, 1995) and recently New Zealand (Kean & Phillips, 2013b). Forty one geographical locations where *Pieris brassicae* is present, some reported as early as 1874 (Dubois & Dubois, 1874), were obtained from Feltwell's (1982) review. The forty one presences were used to construct the historical presences of the species. Overlaying the past presence records with the current occurrence distribution obtained from GBIF (Figure 6.2) revealed that the species still exists in locations with the temperature conditions given in the early publications above. The establishment of the species in these locations is regardless of the fact that the temperatures experienced in these areas are known to be too harsh, for some life stages of the species. Especially for the early instar larvae and the ova. Clearly, life cycle phenology and adaptation is very important in these areas. Because of such a long term

record of the establishment of the species in such areas, it is apparent that the species niche must have a dynamic geographic representation that follows a seasonal gradient as well as a biological adaptation that allows it to survive such conditions.

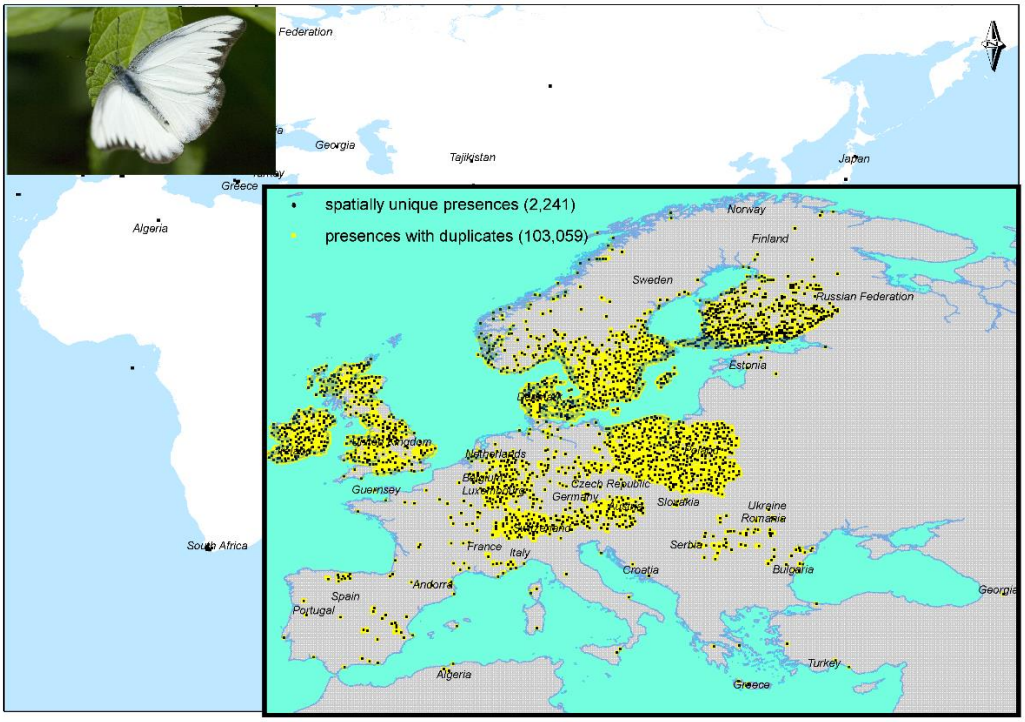


Figure 6.2: Global presence of P. brassicae. *Data courtesy of GBIF*
Yellow areas represent all the 103,059 points accessed from GBIF, black dots represent the 2,241 points used in this study. The New Zealand presence points are not shown here.

6.2.1.4 Biological traits of Pieris brassicae *allowing survival in harsh environmental conditions*

Pieris brassicae has three major mechanisms allowing it to survive harsh environmental conditions which are all partially associated with temperature. The first mechanism is winter diapause (hibernation) and this is employed throughout the *Pieris brassicae* range (Spieth, 2002). However, longer diapause associated with fewer generations are observed in the colder ranges, mostly above northern France (Held & Spieth, 1999). Although, winter diapause is commonly used to escape harsh winter conditions, it is activated by the combined effects of shorter photoperiod with low temperatures (Spieth, 2002). The second survival mechanism used by *Pieris brassicae* is migration and this is usually observed throughout its distribution (Spieth & Cordes, 2012). The third mechanism is summer diapause (aestivation) and this is observed in its warmer ranges of southern Spain, Portugal and Northern Africa (Held & Spieth, 1999). While aestivation by *Pieris brassicae* was first reported by Larsen (1974) in Lebanon, there are no subsequent publications or research that

shows that this is the case for the population in Lebanon. There has been no conclusive research that shows why this adaptation is necessary for the populations in Spain and Portugal except the theory that this adaptation may have been sourced from the North African *Pieris brassicae* populations (Hubert Spieth, personal communication October 10, 2013), and even then it does not explain why the Spanish/Portuguese populations developed aestivation as opposed to the widely used method, migration, to escape high temperatures in its other warm climate ranges like in Israel and Iraq. According to a phenological modelling study carried out by Kean and Phillips (2013d), there is a possibility that the newly introduced *P. brassicae* population in Nelson, New Zealand also aestivates. However, further field study is needed to confirm this result.

6.2.1.5 Constructing the seasonal niche of P. brassicae

By considering the historical information on the realized distribution of the species (Sec-6.2.1.3) together with, the survival mechanisms *Pieris brassicae* uses to survive in its varied environmental range (Sec-6.2.1.4) one can make two assumptions. First, the most vulnerable life stages of *Pieris brassicae* avoid the coldest temperatures that are realized in their geographical range due to winter diapause. Therefore, using the lowest lethal temperature recorded for the pupal stage will identify the fundamental niche of *Pieris brassicae* without risk of overestimating the niche boundary. Second, vulnerable life stages avoid the hottest temperatures in their range by either migration or sometimes aestivation. Therefore, the highest upper lethal temperature will allow the higher temperature bounds of the fundamental niche to be characterised. Using these extreme thresholds on record allows characterisation of seasonal niches of *P. brassicae* even if they migrate or stay in diapause and therefore not always occupy it.

Accordingly, the upper and lower lethal temperature thresholds (Table 6.1) were used to model the *Pieris brassicae* fundamental niche by relating them to two major seasonal temperature events. These are the minimum and maximum temperature of the coldest month and the minimum and maximum temperature of the warmest month. The temperature within the upper and lower lethal temperature of a species was used to construct fundamental thermal niche of the species (Monahan, 2009; Monahan & Tingley, 2012).

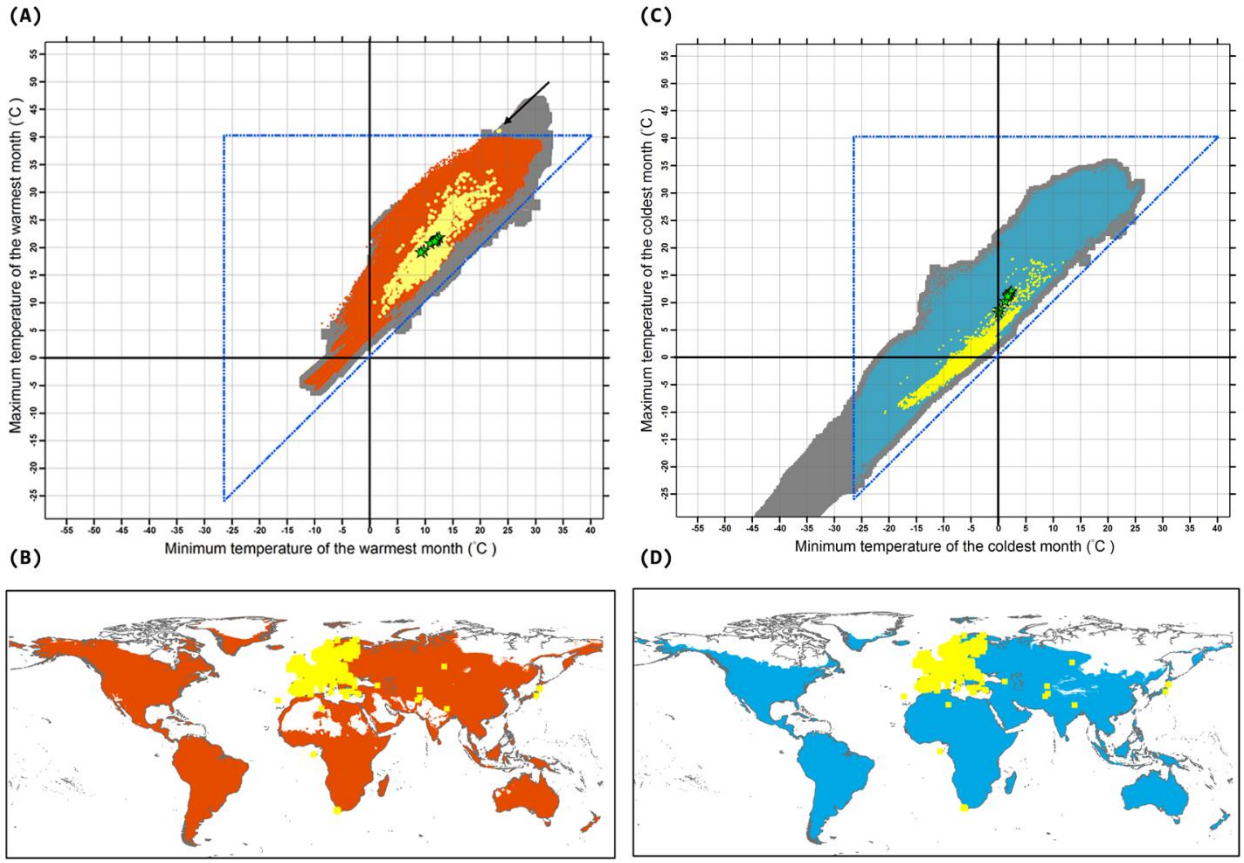


Figure 6.3: The seasonal thermal potential niche of Pieris brassicae *bounded by the upper and lower lethal temperature thresholds and its geographic projection*
(A) and (C) show the potential thermal niche of Pieris brassicae *in the warmest and coldest months in a thermal feature space. Green stars show the recently invaded New Zealand locations in the thermal feature space. (B) and (D) show the geographic projection of the potential niche bounded by the lethal temperature thresholds in the warmest and coldest months respectively. Yellow dots show the global occurrence data. Black arrow points to occurrences recorded outside the lethal thermal thresholds. Note: geographically referenced locations for the well-established populations in Chile and the recent invaded region in New Zealand are not provided in the occurrence dataset, hence are not shown here.*

However, not all temperatures within the upper and lower limit of the species exist in the climate space where the species occurs. Therefore, temperature thresholds obtained from the literature were intersected with the climatic space available in the study area to identify parts of the fundamental niche that are found in the current climate space. This portion of the fundamental niche is known as the potential niche (Monahan, 2009).

Two different realizations of the potential niche were obtained (Figure 6.3) by intersecting the temperature thresholds with two extreme temperature events manifested in the *Pieris brassicae* range, the warmest and coldest months of the year. This demonstrates the dynamic

nature of *Pieris brassicae* niche that is alternatively occupied by migrating populations and/or diapausing populations.

6.2.1.6 Accounting for mobility to unify the seasonal niches of P. brassicae

The seasonally variant suitable areas for the warmest and coldest months respectively (from Figure 6.3-B&D) were overlaid to identify the disconnection between the seasonal niches. This approach was proposed in order to eliminate any identified suitable areas that are only suitable in either the hottest or coldest month and are not close enough to suitable niches for the *Pieris brassicae* populations to immigrate. The distance allowed between the two seasonally suitable areas was 200 km. This distance was chosen by taking the minimum of the distance range (200 - 400 km) reported for migrating *Pieris brassicae* based on a 14-day life span (Spieth & Cordes, 2012). Accordingly, all areas either identified as suitable during both the coldest or warmest months, or areas that are only suitable during the coldest or warmest months but are within 200 km of a suitable area were selected as a part of *Pieris brassicae* potential niche (Figure 6.4).

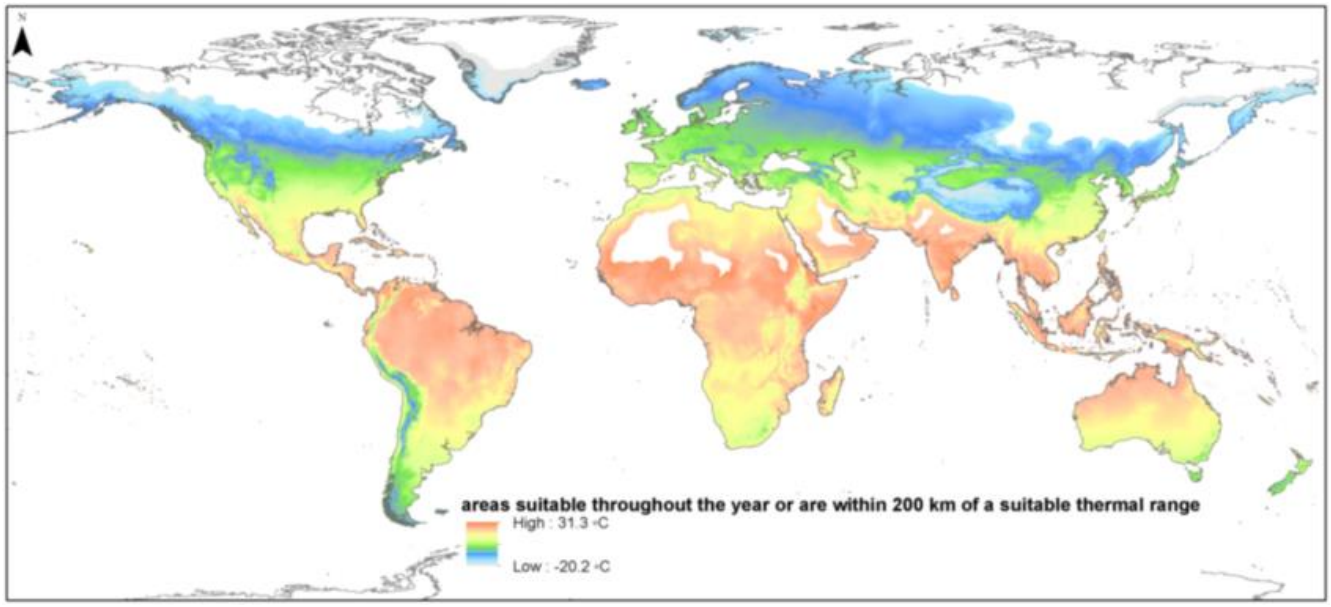


Figure 6.4: The global thermal potential niche of Pieris brassicae.

6.2.1.7 Characterising the fundamental and potential niche of Pieris brassicae

The fundamental and potential niches and other relevant thermal ranges found from the literature were defined (Table 6.3) on the thermal feature space constructed from two computed thermal variables. The variables were temperatures of the maximum warmest month and the minimum coldest month. These variables were capped according to the

seasonal thermal suitability and mobility assumptions incorporated in 6.2.1.5 and 6.2.1.6, to characterise the potential niche based on areas that are either suitable for *P. brassicae* or allow it to persist through diapause or migration. Accounting for migration was achieved by including seasonally suitable areas within 200 km of suitable areas (The "all year" area obtained from Figure 6.4).

Table 6.3: Definitions of the various thermal niche fractions calculated for *P. brassicae*

Name	**Abbreviation**	**Definition**
FN	Fundamental niche	A hypothetical thermal space defined by the upper and lower lethal thermal thresholds of *P. brassicae.* (Table 6.1)
PN	Potential niche	A portion of the fundamental niche (FN) that is available in the current climate space. Additional consideration of mobility of the species is incorporated in defining the potential niche in this study by matching areas within a geographical range of 200 km to a suitable area so that the potential niche is made of the thermal range in which *P. brassicae* could persist, either because it is suitable (within FN) or because it is within a range of 200 k, of a suitable thermal range, for the seasons temperatures are unsuitable for *P. brassicae.*
UFN	Unutilized fundamental thermal niche	Thermal combinations that are not realized in the current climate space in the study area (obtained by subtracting the potential niche from the fundamental niche)
UCCS	Unutilized current climate space	Areas that are unsuitable to *P. brassicae* due to temperatures outside the lethal limits of *P. brassicae* (outside of the FN&PN)
RD'	Projected realized distribution	The projection of the thermal values extracted at the presence points of *P. brassicae* on the thermal feature space, used to estimate the unfilled potential niche of *P. brassicae* with regard to its current locations.
RTR	Realized thermal range	The projection of the isotherms within which *P. brassicae* was reported to occur in early literature. (Recovered from the review of *P. brassicae* natural history by Feltwell (1982))
HSTR	Highly suitable thermal range	A thermal range reported to be optimal for maximum *P. brassicae* generation in a year as well as the minimum time spent on diapause. (Table 6.1)

The relative areas of the various niches in the thermal feature space were calculated. The areas were later used to assess how many of the predicted presences from the correlative model fall within the thermal niche of *P .brassicae*.

The area of the fundamental niche was calculated with a precision of 0.1 ◦C. The total fundamental niche space in ◦C² is calculated using *eq* 6.1.

$$FN = (ULt - LLt)^2 \times 0.5 \quad eq\ 6.1$$

Where ULt = Upper lethal temperature & LLt = lower lethal temperature

The area of the PN, RD', RTR and HSTR were obtained by multiplying the number of cells that fall within each category by the cell resolution of the thermal dataset which is 0.1 ◦C.

Niche fractions were calculated by dividing the various niche types (Table 6.3) by the potential thermal niche (PN) of *P. brassicae*.

6.2.1.8 Deriving physiological suitability surface from the potential niche

While it is difficult to accurately characterize a species niche with only thermal data, it is possible to assume that a species will fill the optimum thermal range before occupying the marginal thermal range in their potential niche, as long as there is no physical barrier that prevents the species from doing so (Hutchinson, 1957; Araújo & Pearson, 2005). Accordingly, the geographic projection of the potential niche constructed by the upper and lower lethal thermal thresholds of the warmest and coldest months was rescaled into a suitability layer with values ranging between zero and one. The most optimum temperature (15 ◦C - Koehler & Geisenhoffer, 2012) found to be the most conducive to *P. brassicae* was set to 1, while all other values that were greater or lesser than the optimum value were rescaled towards 0.

6.2.2 Correlative species distribution modelling

6.2.2.1 Biotic data

Global occurrence points for *Pieris brassicae* were accessed from the GBIF[30] (www.gbif.net) database. There were 103,059 available geo-referenced data points, however, only 2,241 were spatially unique with respect to the resolution of the predictor datasets available. Additional 1,079 occurrence points from the newly introduced New Zealand range were obtained (Craig Phillips, pers. Comm., 23, September 2013). From the New Zealand presences, only two presences were spatially unique for the 10 arc minute resolution of the environmental dataset used for species distribution modelling. The total number of spatially unique presence points used in this analysis was 2,243 points. Four different potential distribution predictions were made according to the following specifications.

The first three predictions were based on three presence datasets that account for varying historical geographic distribution of *Pieris brassicae*. These were used to investigate if the predictive power of the correlative models increase as additional information about the

[30] The list of institutions that provided the occurrence points is given in the appendix 7.1

range of the species is incorporated (Tale 6.4). The fourth[31] prediction was a combination of two sub-predictions that were based on presences from the aestivating and non-aestivating populations of *P. brassicae* respectively (Table 6.4).

Table 6.4 Presence data specification for the correlative models

Correlative predictions	**Code**	**No. of presences used**	**Presence data classification and definition**
1	pbP	2233	Prediction based on Palearctic presences
2	pbPS	2241	Prediction based on Palaearctic + South Africa presences
3	pbPSN	2243	Prediction based on Palaearctic + South Africa + New Zealand presences
4*	pbPSNbt	- -	Prediction based on the complete presence data from PbPSN (n=2243), but first independently modelled according to aestivating and non-aestivating populations of *P. brassicae* and combined to produce a mixed model prediction (described in Chapter 5)
- - 4.1	$pbPSN_{Aes}$	35	Prediction based on aestivating populations of *P. brassicae*
- - 4.2	$pbPSN_{Naes}$	2208	Prediction based on non-aestivating populations of *P. brassicae*

**The fourth correlative model prediction was made using two subsets of the pbPSN presence dataset. The first subset ($pbPSN_{Aes}$, n= 35) included presences from the aestivating population in Spain and Portugal only, while the second subset ($pbPSN_{Naes}$, n= 2,208) contained the remaining presence points from the non-aestivating populations. These two subset predictions were combined to produce a fourth potential species distribution that considered the intraspecific environmental variation in the P. brassicae presence data (Discussed in Chapter 5). The combined prediction was labelled as P. brassicae-Palaearctic-South Africa – New Zealand with biological trait variation (pbPSNbt, n= 2,243). Since the procedure of how to combine such subset predictions was described in chapter 5, only the prediction output of pbPSNbt compared with the other correlative model scenarios is discussed in this chapter.*

Five sets of pseudo-absence points were generated for the three main models and two sub-models defined in Table 6.4 according to the method developed in Chapter 3. The five sets of presence and pseudo-absence points (including subsets of pbPSNbt) were combined to make up the training and test data used to predict potential species distribution of *Pieris brassicae*.

6.2.2.2 Environmental data and model parameterization

The predictors listed Table 6.7 were used to model the potential species distribution of *Pieris brassicae*. The dataset included 35 bioclimatic variables downloaded from the CLIMOND (Kriticos *et al.*, 2012b) website and four topographic variables derived from the SRTM global

[31] The fourth model was used to account for environmental variation within presence data due to biotic traits unique to some populations of a species as discussed in Chapter 5.

digital elevation model dataset (NASA-GSFC, 2000) accessed through the WORLDCLIM data portal (Hijmans *et al.*, 2005b). A random forest classifier was used to select variables.

6.2.2.3 Potential species distribution prediction

The multi-model framework developed by Worner *et al.* (2014) was used to predict the potential distribution of *Pieris brassicae* both at a global and New Zealand extent. Five species distribution models were used for comparison. The models were quadratic discriminant analysis (QDA), Logistic regression (LOG), classification and regression trees (CART), support vector machines (SVM) and artificial neural networks (NNET). Five-fold cross validation with ten repetitions was performed to measure model performance. Model Kappa scores were used to select the best model for each scenario. The predictions from the best models that represent the aestivating ($pbPSN_{aes)}$ and non-aestivating ($pbPSN_{naes)}$ *P. brassicae* population were combined using a map query given in *eq.*6.2. The rule to combine these component predictions was set, so that the prediction from the majority class (non-aestivating presences) was taken whenever a presence is predicted, if not, the prediction from the minor class is assigned. If both clusters did not predict a location, the average of the component predictions was assigned to the combination prediction pbPSNbt.

$$Con\left((\text{PbPSNnaes} >= 0.5), PbPSNnaes, Con\left((PbPSNnaes < 0.5)\ \&\ (PbPSNaes >= 0.5), PbPSaes, \frac{PbPSNnaes + PbPSaes}{2}\right)\right) - - - eq\ 6.2$$

Model predictions across all four scenarios were compared to investigate if there is any convergence in prediction between correlative and mechanistic niche models as more presence information becomes available.

The best correlative model scenario to be used for the hybrid prediction was selected by comparing the percentage of correctly predicted presences for the target study area, New Zealand, within the potential niche characterised by the mechanistic model.

6.2.3 Hybrid model prediction

Hybrid prediction was produced using the potential *Pieris brassicae* distribution prediction from the best correlative model and the physiological suitability layer derived from the generalized mechanistic niche model. The probabilistic suitability predictions from the two

models were combined using a simple "either or" logical classifier specified for this study (Eq. 6.3). The rules followed to hybridize the predictions and their justification are given below.

1. *The correlative model prediction that best filled the potential thermal niche identified by the mechanistic model was selected. An overlay analysis was performed to discard any areas predicted by the correlative model outside the potential thermal range of* P. brassicae.

2. *When the probability of the correlative model is >= 0.5, the hybrid prediction was given the correlative model prediction. The correlative model value was chosen over the mechanistic one, because it was derived from more environmental covariates, which has the advantage of differentiating the level of suitability of isothermal cells over the landscape giving better spatial detail.*

3. *When the probability of the correlative model is < 0.5 but when the mechanistic model > 0.5 the hybrid was assigned the values of the mechanistic prediction. In such cases, the correlative model has not picked up these sites, in spite that they are physiologically suitable. That could be because a representative environmental condition was not provided in the form of occurrence points and that is usually when the mechanistic model complements the correlative model.*

4. *When both predictions were not suitable then the area was confirmed by the two models as unsuitable and was given the average of the two predictions.*

$$Con\left(\text{Corr} \geq 0.5, \text{Corr}, Con\left((\text{Corr} < 0.5)\&(\text{Mech} \geq 0.5), \text{Mech}, \frac{Corr + \text{Mech}}{2}\right)\right) - - - - - - eq6.3$$

Where Corr = the potential suitability prediction from the selected correlative model & Mech = the physiological suitability prediction from the mechanistic model. Con is the function call in ®Spatial Analyst extension of ArcGIS to perform conditional operations on raster data.

6.3 Results and discussion

6.3.1 Physiological suitability - generalized mechanistic niche model prediction

The fundamental niche of *Pieris brassicae* bounded according to its lethal upper and lower temperatures, is shown in the two dimensional thermal feature space of the minimum coldest month against the maximum warmest month (Figure 6.5-A). The geographic projection of the potential niche identified extensive areas (Figure 6.5-B) due to the rather

high upper lethal temperature (ULt) and low lower lethal temperature (LLt) thresholds of *P. brassicae*.

In reality the fundamental niche of any species is defined by an *n* dimensional hyper-volume where *n* stands for the number of variables that physiologically affect the species (Colwell & Rangel, 2009). Although, it is not possible to account for all variables and their interactions that affect a species in the real world, additional variables to temperature should better characterise the potential niche of a species, especially if the species has extreme lethal thresholds.

Table 6.5 *Pieris brassicae* niche fraction estimates based on lethal temperature thresholds.

Fractions of Climate/niche space	**Fraction (%)**	**Remark**
PN/FN	43.76	Note: WORLDCLIM (50 yr. average data) was used to define the current climate space, based on which the potential niche was approximated. ∴ any predictions are likely to show the global long term thermal limitations to *P. brassicae* distribution (Monahan, 2009).
RD'/PN	4.48	
RTR/PN	41.22	
HSTR/PN	1.03	

The projected realized distribution (RD') which was characterised by the current presence points of *P. brassicae* on the thermal feature space, show that *P. brassicae* is not at equilibrium with its thermal niche as there are areas that are not filed by its current presence records (Table 6.5). However, it is difficult to estimate exactly how much of its niche is vacant because not all presence points of *P. brassicae* in reality were accounted for by the GBIF database. The term "vacant" here is hypothetical and only refers to being vacant of *P. brassicae*, as other organisms share this thermal niche. Nevertheless, the fact that the RD' is located in the middle of the PN shows that the species is currently not in areas that are at its thermal limit, meaning it can successfully invade new habitats if other factors like host availability and means of dispersal are fulfilled.

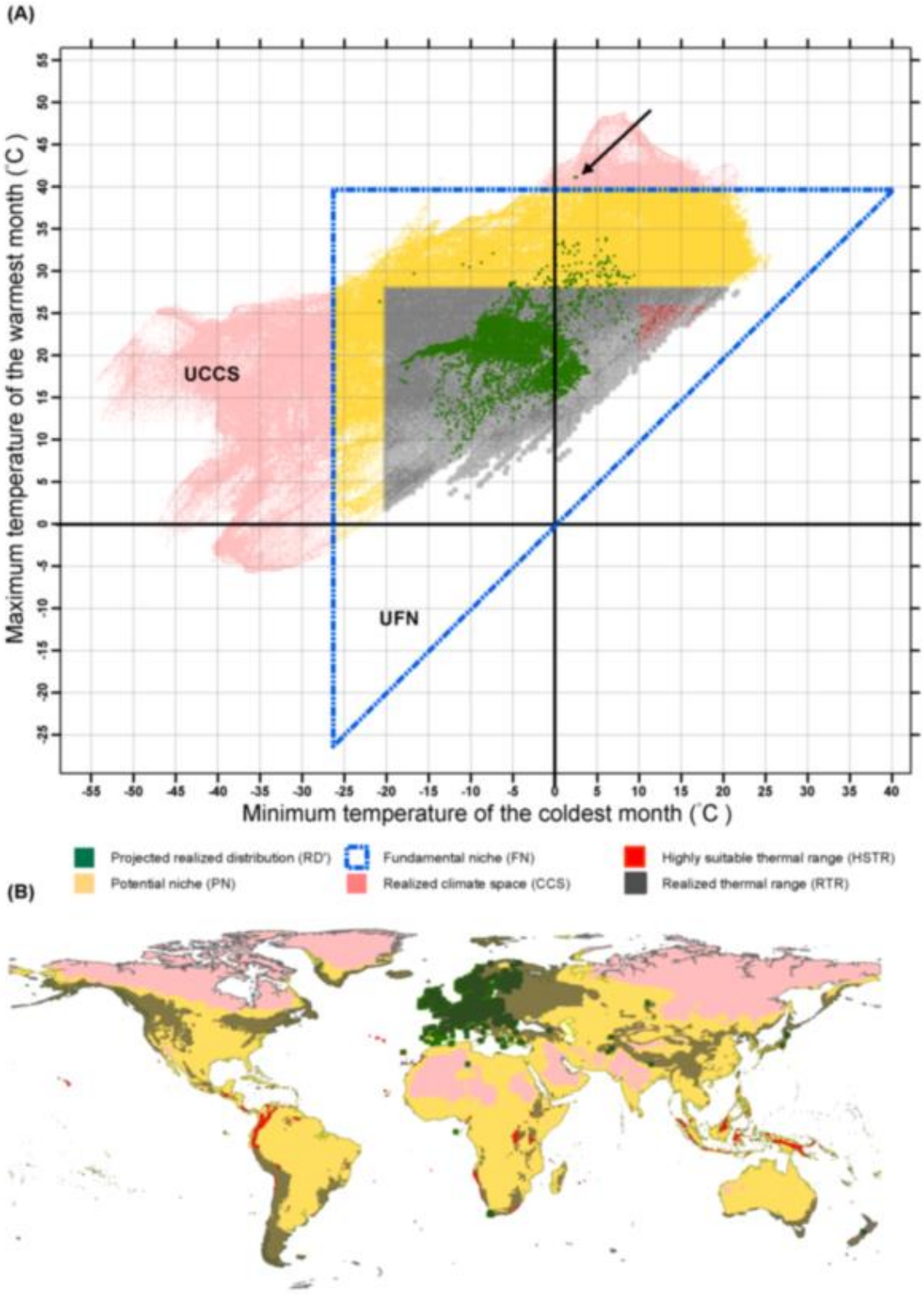


Figure 6.5: The potential niche (PN) and projected realized distribution (RD') of Pieris brassicae *(A) in the thermal feature space (B) in the geographic space*
Arrow shows a presence point found outside the potential thermal niche of the species, the outlier is a presence point recorded in Algeria. This outlier can be the result of inaccuracy in the interpolated WORLDCLIM grid due to inadequate weather stations in the area.

The geographic projection of the thermal range defined as RTR using the mechanistic niche model showed a very good agreement with the native Palearctic distribution of *Pieris brassicae* and has successfully predicted all the introduced ranges of *Pieris brassicae* in Chile, South Africa and New Zealand. All areas in New Zealand were found to be thermally suitable for *P. brassicae*.

The HSTR portions of the potential niche, when projected onto the geographical space, identified areas that have similar thermal conditions as those found optimal in empirical studies (thermal values and reference given in Table 6.1). The geographical projection of the HSTR (Figure 6.5 –B) were far removed from the core population distribution in Europe. The fact that most of these were located close to the tropics seems strange. However, a closer investigation of these locations show that they are all located in high altitude where the temperature is kept between the most suitable ranges (15°C -26°C) all year round. Such conditions remove the need for longer diapauses and could maximize the number of generations *P. brassicae* can have per year (Spieth & Sauer, 1991).

HSTR covers a very small percentage of the global area (0.24%) compared to the potential niche (PN), this highlights an interesting fact that, optimal conditions obtained from empirical experiments like the HSTR can be very rare in the field (Table 6.6). This can create an uncertainty around some mechanistic predictions based on parameters from a laboratory experiment with controlled environmental conditions

Table 6.6 Global coverage of the potential thermal niche and various important thermal ranges estimated for *Pieris brassicae.*

Climate/Niche space	**Variable space (°C²)**	**Global geographical coverage (%)**
Total available thermal current climatic space (CCS)	1492.70	100*
Fundamental niche space (FN)	2204.48	--**
Potential (PN)	964.75	67.90
Projected realized distribution (RD')	43.18	0.38
Realized thermal range (RTR)	397.74	17.53
Highly suitable thermal range (HSTR)	10.02	0.24

** The available climate space represents all thermal climates realized in the world thus 100% of the global area*

*** The fundamental niche space is conceptual and only a portion of it is realized in the current climate space as PN*

Coincidentally, even if photoperiod was not used as a variable to define the HSTR, the relatively low variation in day length in areas identified by the HSTR also mean,

deactivation of facultative diapauses for *P. brassicae*, increasing the number of generations due to less time spent while in diapause. As shown in the map these areas are highly isolated from where the core distribution is and thus the chance of the species getting there might be limited. The availability of host plants is of course a basic criterion for *P. brassicae* to establish in these areas, but considering its host species are the *Cruciferae* family that are grown all over the world for agriculture purposes, host plant presence is unlikely to be a limiting factor to *P. brassicae*. South Africa is the only area from the globally identified HSTR regions where *P. brassicae* is confirmed to be established.

According to the physiological suitability surface (Figure 6.6), the North Island of New Zealand generally looks thermally more suitable than the South Island. However, it is important to note that most of New Zealand is already within the potential niche and the RTR range identified in the thermal feature space. Therefore, the whole country is within *P. brassicae* thermal tolerance except for the north-western tip of the North Island that are not coloured in Figure 6.6, which are unsuitable to *P. brassicae*. The physiological surface in Figure 6.6 shows levels of thermal suitability throughout the country.

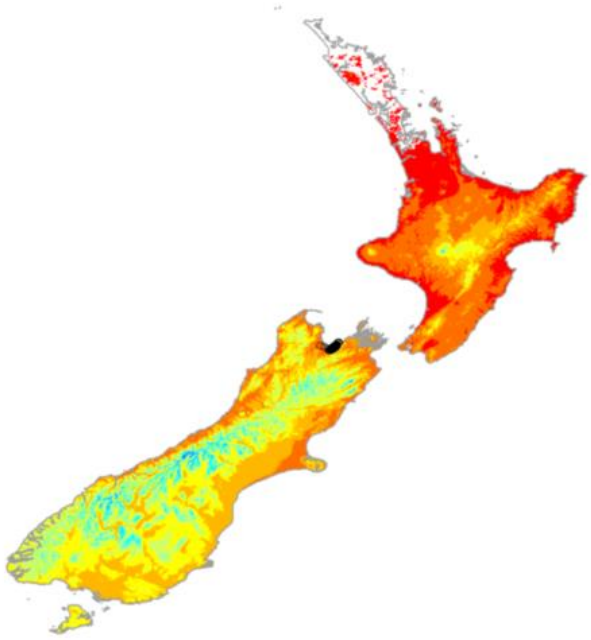


Figure 6.6: The New Zealand extent physiological suitability surface based on thermal thresholds of P. brassicae *derived from the mechanistic niche model.*
The black circle shows the area where P. brassicae *is recently introduced. Warmer shades show higher level of suitability. All areas in New Zealand except the not coloured areas at the tip of the North Island were however identified as thermally suitable for* P. brassicae *as they fall within the realized thermal range.*

6.3.2 Environmental suitability - correlative species distribution model prediction

The variables selected for the first two correlative model scenarios that have the native range presence points (pbP) and additional points from the South African *P. brassicae* population (pbPS) were identical (Table 6.7). That suggests that the introduced South African range does not vary much environmentally from the native range. However, when the third scenario that has presences from the New Zealand population (pbPSN) was compared with the previous two scenarios, different variables were selected. Four of the original variables did not explain the pbPSN dataset as a whole and were omitted, and one new variable that was not originally selected was added (Table 6.7). The discrepancy in variable selection shows that there was a considerable environmental variation introduced from the newly added New Zealand points. The fourth scenario (pbPSNbt) is a combination prediction of two separately modelled subset populations. The variables from pbPSNbt were different from the first three scenarios (Table 6.7).

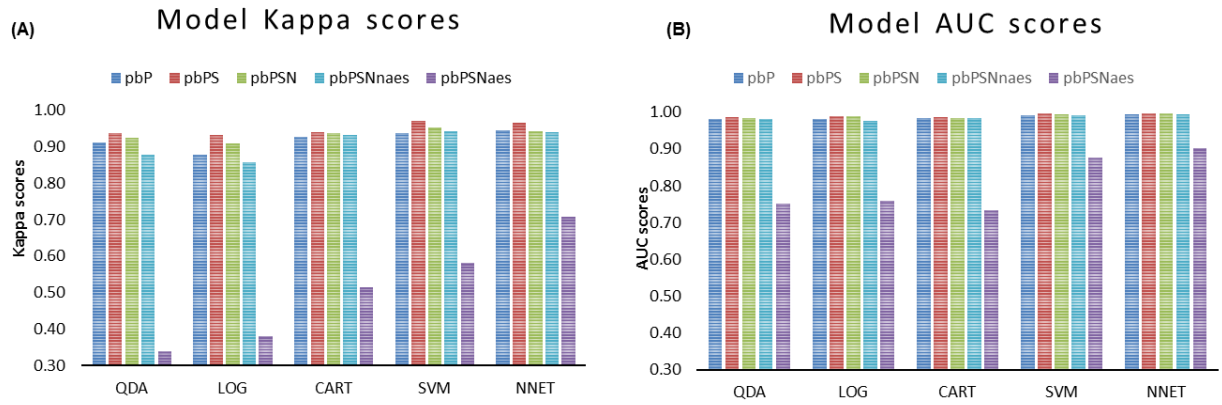


Figure 6.7: Model performance for the four scenarios, (A) Kappa scores, (B) AUC scores, (predictions from pbPSN$_{Naes}$ and pbPSN$_{aes}$ are subsets of the pbPSNbt scenario).

The best model for each scenario was selected using model Kappa scores (Figure 6.7- A). NNET, SVM and SVM were selected for the first three scenarios respectively. The fourth scenario was a combination of two subset predictions; SVM was selected for the non-aestivating population prediction while LOG was selected for the aestivating population prediction. All models performed well (Kappa score > 0.81), except for models based on the aestivating population training dataset. The performance of models using presences from the aestivating population were disadvantaged by the small sample size of the data (n = 70, for both presence and pseudo-absences). The performance of the best model (LOG, Kappa

score = 0.7) for the aestivating population prediction had an acceptable Kappa (Figure 6.7-A). It is important to note that AUC usually gives inflated values for models using small sample size for training (Figure 6.7-B), therefore Kappa scores were used for model selection.

Table 6.7: Variables selected for the four different training datasets

Var. No	Variable Name	pbPA	pbPS	pbPSN	pbPSNbt*
01	Annual mean temperature (°C)	✓	✓	✓	✓
02	Mean diurnal temperature range (mean(period max-min)) (°C)	✓	✓	✓	✓
03	Isothermality (Bio02 ÷ Bio07)				
04	Temperature seasonality (C of V)				
05	Max temperature of warmest week (°C)	✓	✓	✓	✓
06	Min temperature of coldest week (°C)	✓	✓		
07	Temperature annual range (Bio05-Bio06) (°C)	✓	✓	✓	
08	Mean temperature of wettest quarter (°C)				
09	Mean temperature of driest quarter (°C)	✓	✓		✓
10	Mean temperature of warmest quarter (°C)	✓	✓	✓	✓
11	Mean temperature of coldest quarter (°C)	✓	✓	✓	
12	Annual precipitation (mm)				
13	Precipitation of wettest week (mm)				
14	Precipitation of driest week (mm)	✓	✓	✓	✓
15	Precipitation seasonality (C of V)	✓	✓		
16	Precipitation of wettest quarter (mm)				
17	Precipitation of driest quarter (mm)	✓	✓	✓	✓
18	Precipitation of warmest quarter (mm)				
19	Precipitation of coldest quarter (mm)				
20	Annual mean radiation (W m^{-2})	✓	✓	✓	✓✓
21	Highest weekly radiation (W m^{-2})				
22	Lowest weekly radiation (W m^{-2}	✓	✓	✓	✓
23	Radiation seasonality (C of V)	✓	✓	✓	✓
24	Radiation of wettest quarter (W m^{-2})				
25	Radiation of driest quarter (W m^{-2})				✓
26	Radiation of warmest quarter (W m^{-2})	✓	✓		✓
27	Radiation of coldest quarter (W m^{-2})	✓	✓	✓	✓
28	Annual mean moisture index	✓	✓	✓	✓
29	Highest weekly moisture index				
30	Lowest weekly moisture index	✓	✓	✓	✓
31	Moisture index seasonality (C of V)				
32	Mean moisture index of wettest quarter				
33	Mean moisture index of driest quarter	✓	✓	✓	✓
34	Mean moisture index of warmest quarter				
35	Mean moisture index of coldest quarter			✓	✓
36	Elevation (m)				
37	Slope (deg)				
38	Aspect (deg)				
39	Hillshade				

** In the pbPSNbt column red ticks show variables chosen for the aestivating population model ($pbPSN_{aes}$) and black ticks show variables chosen for the non-aestivating model ($pbPS_{Naes}$)*

The effect of the different training datasets is also shown by the predictions for the four scenarios (Figure 6.8). The total predicted suitable area has generally increased as more presence points from newly introduced ranges were used in the training dataset (Table 6.8).

The predictions from all scenarios agree with respect to the native distribution of *P. brassicae*. The area of predicted presences increased in the last three scenarios (pbPS, pbPSN & pbPSNbt) when compared with the first scenario (pbP) that had presence points only from the native range of *Pieris brassicae.* This result is supported by other studies that concluded using presences from the introduced range of invasive species, gives high predictive power to correlative models (Urban *et al.*, 2007; Beaumont *et al.*, 2009; Rodda *et al.*, 2011).

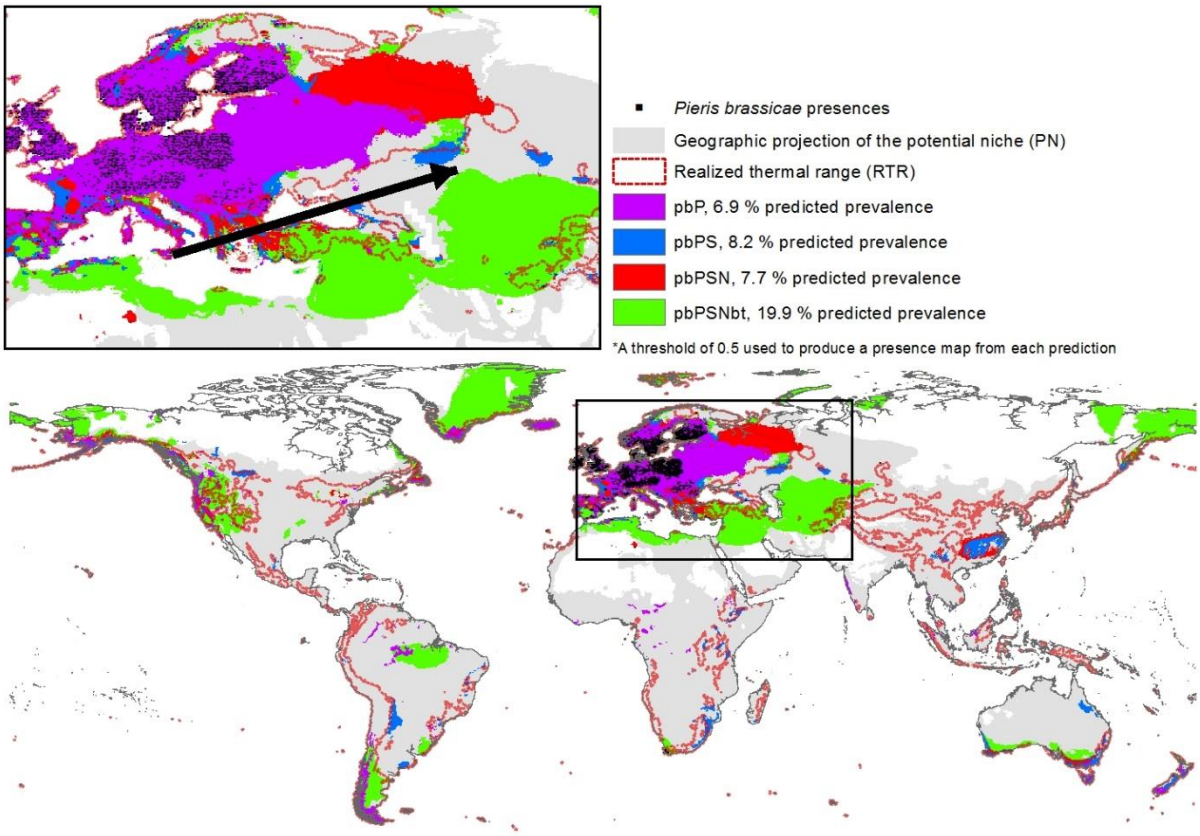


Figure 6.8: The global potential P. brassicae *distribution for the different training dataset scenarios. The boundary of the realized thermal range (RTR) identified by the mechanistic niche model is overlaid here for comparison. The black arrow shows the direction of increasing predictions in the geographic range for the four scenarios given in the legend.*

The high resolution (30') New Zealand extent potential *P. brassicae* distributions projected from the best models for the global prediction of the four scenarios were compared. The first scenario pbP, with a training data that only represents the typical Palearctic distribution of the species completely missed the newly introduced range in New Zealand (Figure 6.9-A). The second dataset pbPS had an extra eight presence points from the introduced range in South Africa. This prediction gave the higher predicted prevalence than pbP. However it did not identify the newly introduced population in New Zealand (Figure 6.9-B).

The third dataset pbPSN, that included all available presence points including the presences in New Zealand, as expected predicted the core established population in New Zealand (Figure 6.9-C). The last dataset, pbPSNbt, which was the combination of the subset predictions for the aestivating and non-aestivating *P. brassicae* populations, gave similar predictions with pbPSN but identified more suitable areas in the North Island of New Zealand (Figure 6.9-D).

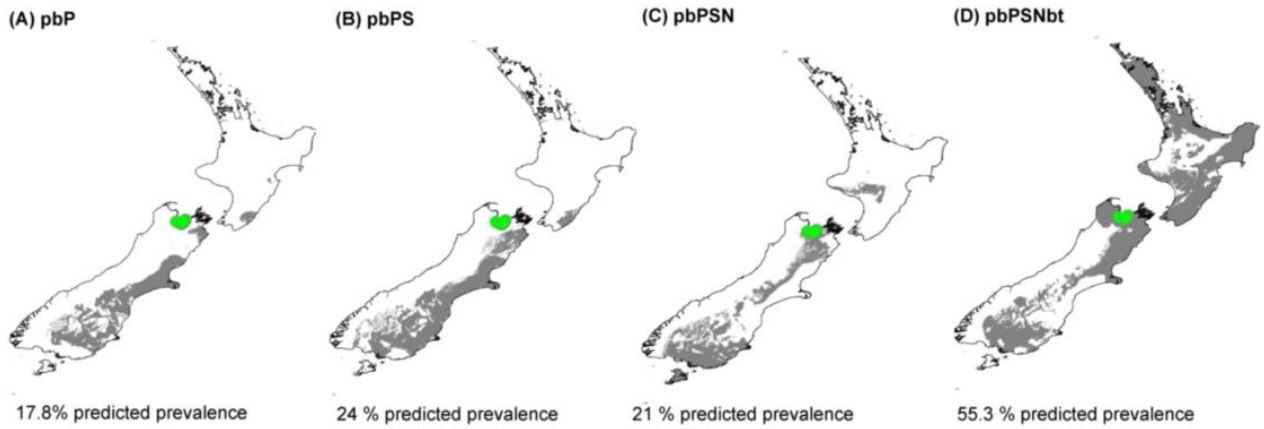


Figure 6.9: The potential P. brassicae *distribution predicted for New Zealand, according to the four varying training data scenarios. The green circles show* P. brassicae *presences in New Zealand.*

6.3.3 Hybrid potential species distribution prediction

6.3.3.1 Comparison between correlative and mechanistic predictions

Predictions that are common among correlative models were within the realized thermal range (RTR) that was reported in the literature about *P. brassicae* distribution in the 80's (Figure 6.8). The filling of the potential niche by the correlative model predictions generally increased as more presences from newly invaded areas were added to the presence datasets.

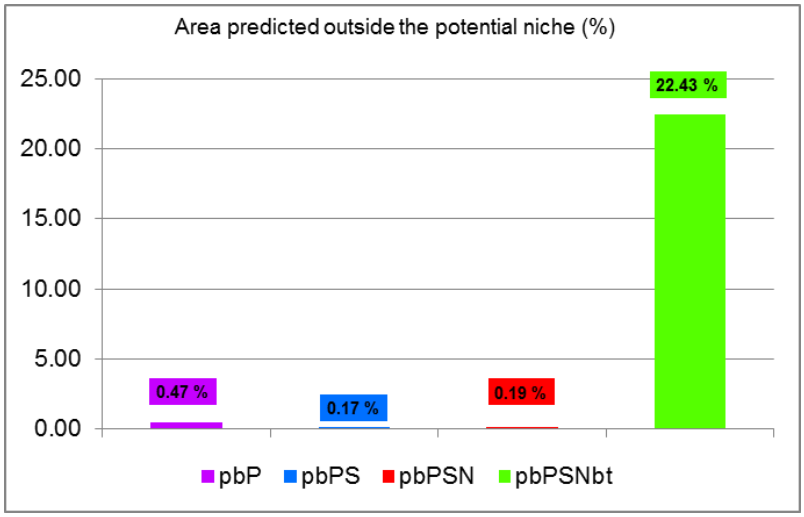


Figure 6.10: Percentage of predicted presences that are outside the potential niche (PN) by the different correlative model scenarios Based on geographical areas calculated in decimal degrees ($^{\circ 2}$)

The correlative model predictions that were outside the potential thermal niche identified by the generalized mechanistic model were considered over-predictions. The first three scenarios (pbP, pbPS & pbPSN) were all well within the potential niche, whereas the fourth scenario (pbPSNbt) which was a combination of two subset predictions, identified areas that are outside the potential niche (Figure 6.10 & Figure 6.11).

The geographic locations of the identified over-predictions of the four correlative models is given in Figure 6.11, the geographical boundary of the potential niche is also shown for comparison.

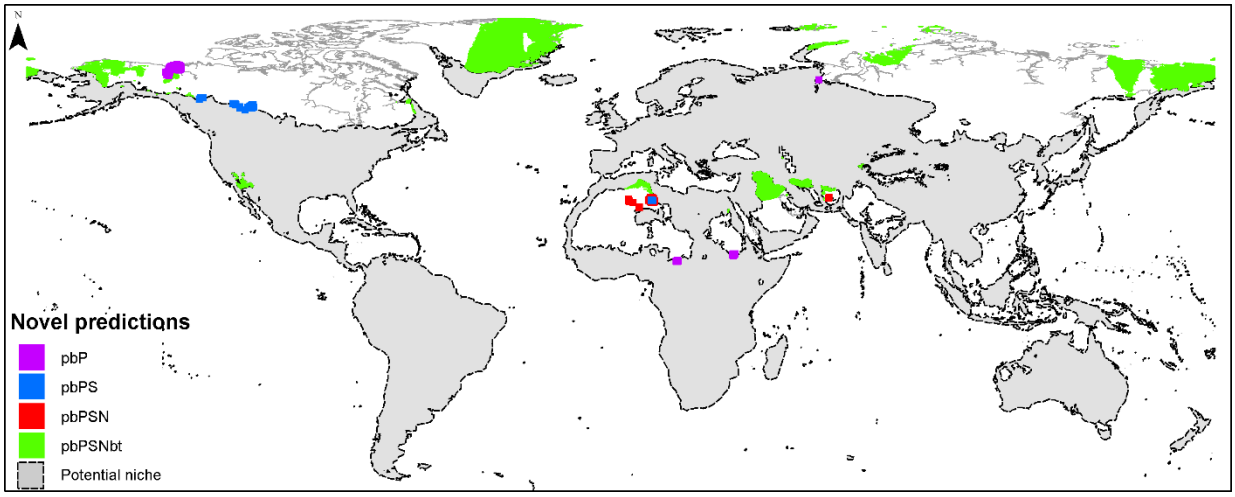


Figure 6.11: Areas of over-predicted thermal ranges identified as suitable by the four correlative predictions.

The over-predicted areas were removed from the correlative predictions as they were outside the thermal limits of *P. brassicae*. However, one cannot be completely be sure about labelling the areas that fell outside the potential niche projection as "over-predicted". It is also possible that these areas are included in the different snapshots of the potential niche from the varying seasons because the geographical realization of the potential niche is dynamic depending on seasonal changes (Tingley *et al.*, 2009). Therefore, it is important to investigate a number of annual thermal events and their geographical projection to see if these novel areas are not included in any timeline of the mechanistic niche model predictions.

With respect to the correlative models, additional to the error caused from internal model calibrations, inclusion of seasonal presences in correlative models could create such over-predictions. Some areas where *P. brassicae* is recorded might be from areas where transient

populations are found. For example, there are a number of presences in the GBIF record that show *P. brassicae* presence at the northern coast of Norway, where *P. brassicae* is present only during the warm months (Feltwell, 1982). Because of the static nature of most correlative SDMs, their predictions lack the power to distinguish between areas where the species is always present and those it occupies only certain months of a year (Franklin, 2010b). A possible method to deal with this staticity is implementing the use of stacked species distribution models (s-SDMs) (Calabrese *et al.*, 2014). Previous studies have proposed stacking SDMs to estimate species richness and to carry out community level niche modelling. However, it is possible to use s-SDMS to produce a dynamic potential species distribution by modelling monthly suitability of an area based on monthly climatic data and matching presences, and stacking the predictions for 12 months of the year. Such tracking of habitat suitability throughout the year will provide an improved representation of potential distribution of species and reduce under- or over-predictions from SDMs.

From the correlative scenarios, pbPSNbt was the model that filled the potential niche and other identified important thermal ranges of *P. brassicae* the most (Table 6.8). Therefore it was selected as the best correlative model to proceed with the hybrid correlative-mechanistic prediction.

Table 6.8: Areas predicted by the four correlative models in comparison with the mechanistic niche characterizations

Scenario/ best model	Tot. predicted area ($^{\circ 2}$)	% of PN	% of RTR	% of HSTR
pbP/NNET	1,126.9	9.78	44.36	7.48
pbPS/SVM	1,342.6	11.68	41.72	0.50
pbPSN/SVM	1,272.6	11.07	41.02	0.72
pbPSNbt/LOG&SVM	2,550.5	17.15	46.99	9.06

The fact that the best correlative model to fill the potential niche the most is also the model that over-predicted the most warrants some discussion. The fourth scenario (pbPSNbt) combined predictions from the two sub-predictions of models trained on the aestivating and non-aestivating populations of *P. brassicae*. The rule used to combine these two predictions was set such that maximum sensitivity is obtained from the combined prediction (Eq. 6.2). A choice of maximizing sensitivity, specificity or optimizing between the two is a decision that depends on the underlying objective of the study. Optimizing between sensitivity and specificity is usually undertaken to calibrate models that are used for further projections

using independent data. Also it is the logical choice when either sensitivity or specificity are not explicitly required due to the objective of the research (Pearson *et al.*, 2004) as discussed below.

Maximizing specificity reduces error of commission and is desirable to identify areas for conservation and especially for relocation of endangered species. Prediction for conservation studies need to be specific due to the pressure to prioritize conservation areas either because of budget constraints or increasing demand of land for productive industries. Whereas maximizing sensitivity reduces error of omission, and mostly required when predicting potential species distribution for invasive species (Araújo & Peterson, 2012). The latter was appropriate for this study because the aim was to identify areas suitable to *P. brassicae* in its newly invaded ranges, especially in New Zealand.

Model pbPSNbt identified most of the areas within the PN (Table 6.8) and also correctly predicted potential suitable areas in New Zealand (Figure 6.9) as identified by the generalized mechanistic model. The increased predictive power from individually accounting for variation in the presence data was clearly shown by the filling of the potential niche by this model more than the predictions from the other three models. The maximized sensitivity in the pbPSNbt prediction resulted from the way its two sub-predictions are combined, which was at the cost of minimized specificity that resulted in predictions outside the geographic projection of the potential niche of *P. brassicae*. Referring to the mechanistic model outputs in this case was beneficial, as it enabled the discrimination of potential false positives in the prediction of the pbPSNbt model. The pbPSNbt prediction was therefore corrected by removing predicted areas that fell outside the geographic projection of the potential niche identified by the generalized mechanistic niche model.

The most straightforward way to benefit from the predictions of correlative and mechanistic models is to compare the outputs (Kearney *et al.*, 2008) as done here. The level of consensus between the correlative and the mechanistic predictions, could be used to test a number of hypotheses about the species distribution. For example, agreement in prediction shows high probability of accuracy of the potential distribution for the species. It is important to note however, that the species might not actually be present even in sites predicted by both mechanistic and correlative models, as the realized distribution of a species is affected by

more than abiotic factors. Nevertheless, agreement between the mechanistic and the correlative model gives higher confidence in the predicted distribution. At least, agreement between the two predictions shows that the area predicted by the correlative model is within the physiological limit of the species.

Disagreement between predictions however can have more than one explanation. When the correlative model falls short of filling the areas identified as physiologically suitable by the mechanistic model it could be because representative presence points were not available to cover the area. Or if the mechanistic model did not cover areas predicted by correlative models, it could either be the biological parameters used to calibrate the mechanistic model do not account for complex environmental interactions that might alter the conditions set for suitability, or the correlative model has over-predicted. Usually the latter is assumed when there is a disagreement between mechanistic and correlative models, however this has to be determined case by case depending on how well the mechanistic model was parameterized and whether the correlative model has extrapolated.

6.3.3.2 Correlative-mechanistic prediction

A hybrid map was produced by combining the corrected pbPSNbt prediction with the mechanistic physiological suitability surface. The correlative model typically predicts the habitat suitability of an area for a species based on occurrence data and the associated underlying environmental variables (Sillero, 2011). On the other hand, the physiological suitability surface from the mechanistic model can be used to identify areas that are within the thermal tolerance of the species even if these areas were never occupied before. Hence, complementing areas the correlative model might have missed simply because no occurrence data was provided from these suitable ranges.

The potential *P. brassicae* distribution for New Zealand, according to the pbPSNbt prediction (correlative model), the physiological suitability surface (mechanistic model) and their hybrid prediction is given in Figure 6.12 - A, B and C respectively. The most important observation was that the correlative model did not identify all locations that were suitable for *P. brassicae* in the North Island of New Zealand based on thermal tolerance. It was apparent from Figure 6.9 that the predictive power of the correlative models increased with the increasing information on the environmental range of *P. brassicae*, as the area predicted

in New Zealand kept on increasing with every model that used more presence points from outside the native range of *P. brassicae*. Because the mechanistic model showed New Zealand as generally suitable for *P. brassicae*, these increasing predictions from the correlative models are less likely to be over-predictions. Considering that *P. brassicae* is still invading new habitats, it is difficult to summarise the correlative model chosen in this study also had the complete information on *P. brassicae* environmental range. Therefore, the additional areas identified as suitable by the hybrid prediction are especially informative.

A hybrid prediction can be more informative than the individual correlative and mechanistic models. For example, the areas identified as highly suitable by the mechanistic model ($p > 0.9$) in the north-eastern part of the North Island of New Zealand (Figure 6.12-B), had reduced suitability in the hybrid prediction (Figure 6.12-C). From investigation of the three moisture based variables selected by the pbPSNbt model, it was apparent that this area was too dry for *P. brassicae*. Therefore, the additional information from the multiple environmental variables used by the correlative model was useful in providing spatial detail to the hybrid prediction.

The hybrid global prediction incorporated areas that were not predicted by the correlative model. For example, an extensive area in the eastern North America that was not predicted by the correlative model (Figure 6.12-D) was predicted as thermally suitable for *P. brassicae* (Figure 6.12-F).

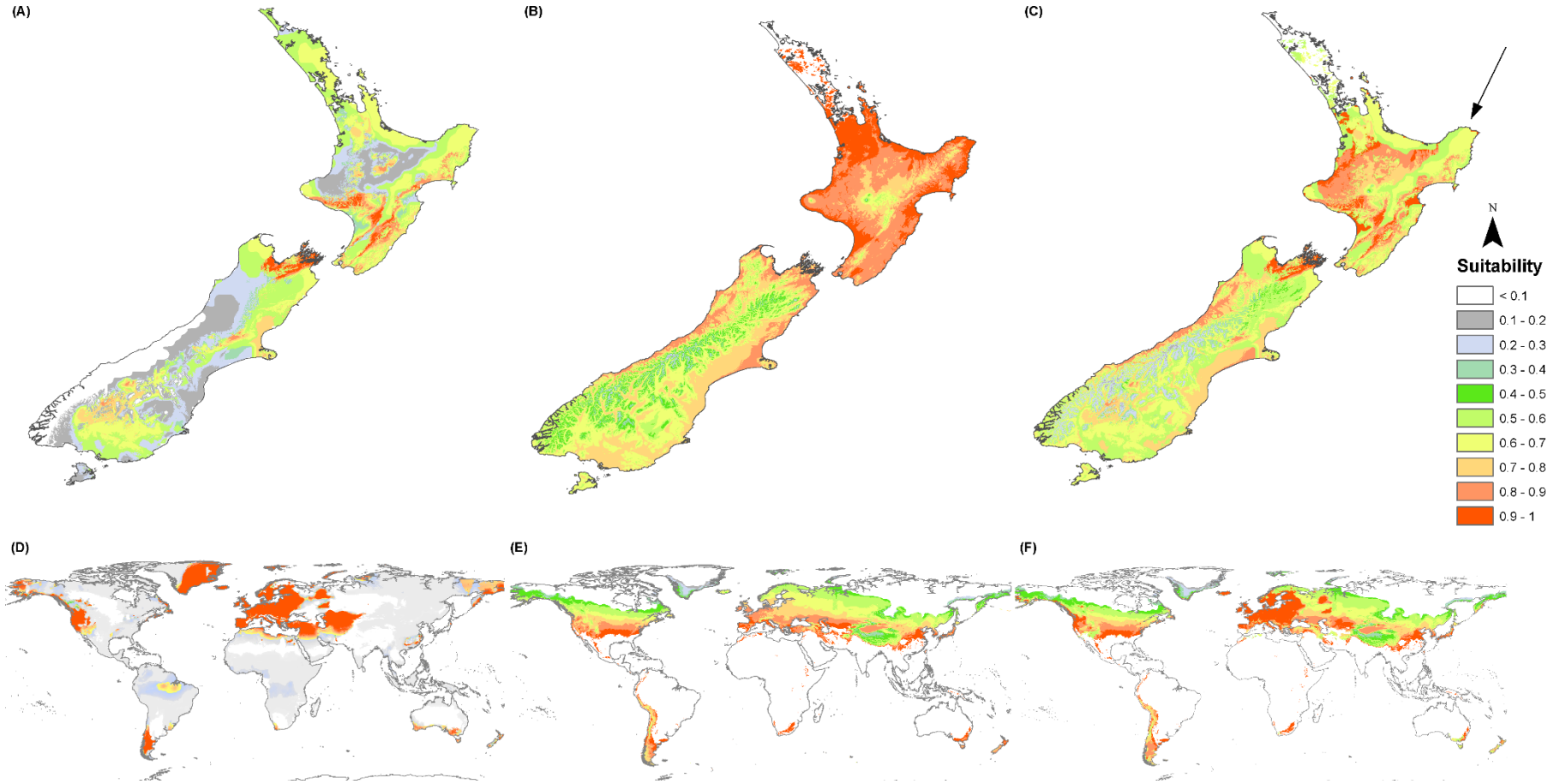


Figure 6.12: Comparison between global and New Zealand extent predictions of the correlative, mechanistic and hybrid models

Potential distribution of P. brassicae *in New Zealand as illustrated by the (A) Correlative distribution prediction (B) the physiological thermal suitability surface and (C) the hybrid prediction. (D), (E) and (F) give the global distribution in the same order. The arrow shows areas that were found thermally highly suitable but were given less overall suitability because the correlative model identified that other important variables are less favourable in that location. That area was found to be too dry for* P. brassicae *even if the temperature was suitable.*

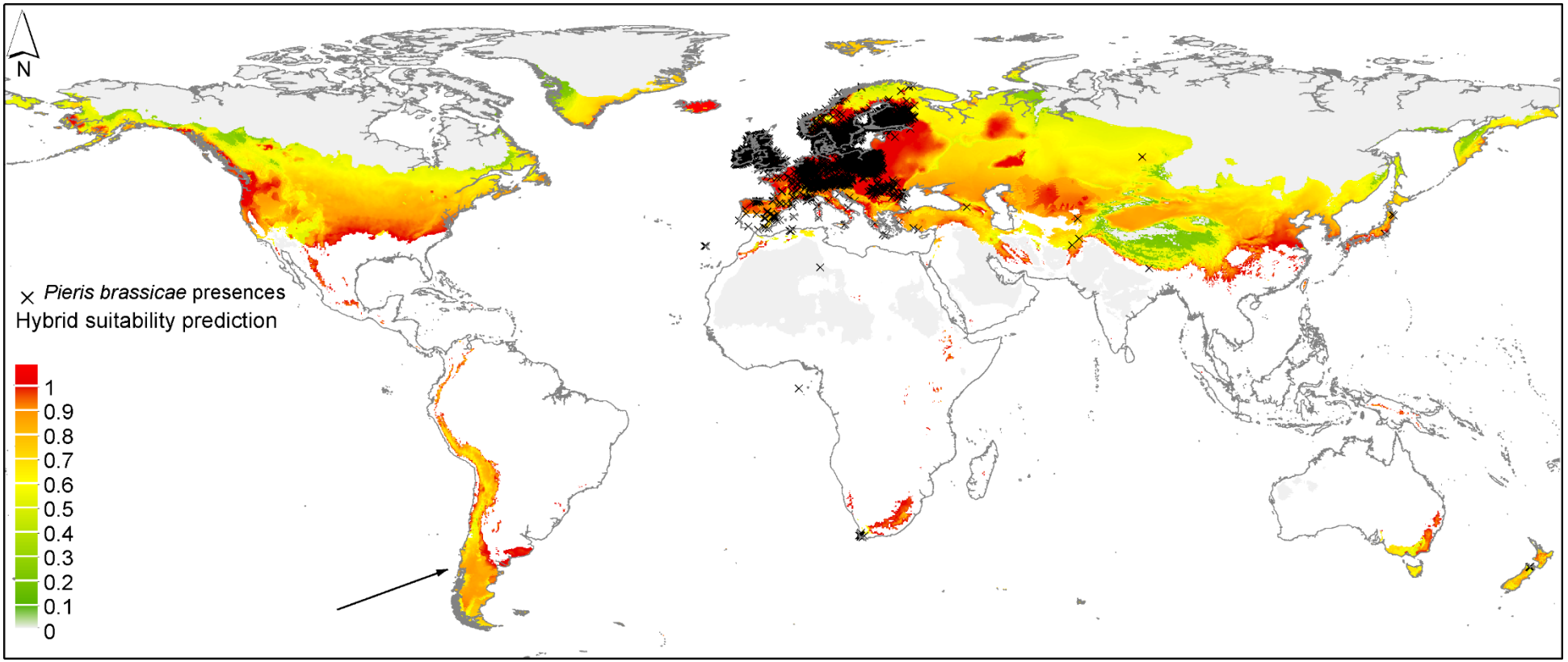


Figure 6.13: Global hybrid potential P. brassicae *distribution prediction*

Enlarged map of the global hybrid prediction of the potential distribution of P. brassicae. *Novel predictions found outside the potential niche of the species are discarded to avoid overestimation of the geographical distribution of* P. brassicae.

The arrow points to the general location where P. brassicae *has established in S. America, presence points from this area in Chile were not used both in model training or testing as there were no geographically referenced data available to this study.*

6.4.3.3 Implications on the future dispersal of Pieris brassicae *in New Zealand*

According to thermal requirement of *P. brassicae* almost all of New Zealand is within the RTR identified by the mechanistic niche model. Therefore, it is possible to conclude that the species will not encounter any thermal zones that are outside its tolerance in New Zealand and we can eliminate temperature gradient as deterrent to its future dispersal within New Zealand.

The hybrid prediction showed that the western side of the North Island and the eastern side of the South Island are more suitable than the rest of the country. Nelson city, the area where *P. brassicae* was first detected in 2010 has a generally lower thermal suitability than the surrounding districts like Tasman and Marlborough where considerable highly suitable areas were identified.

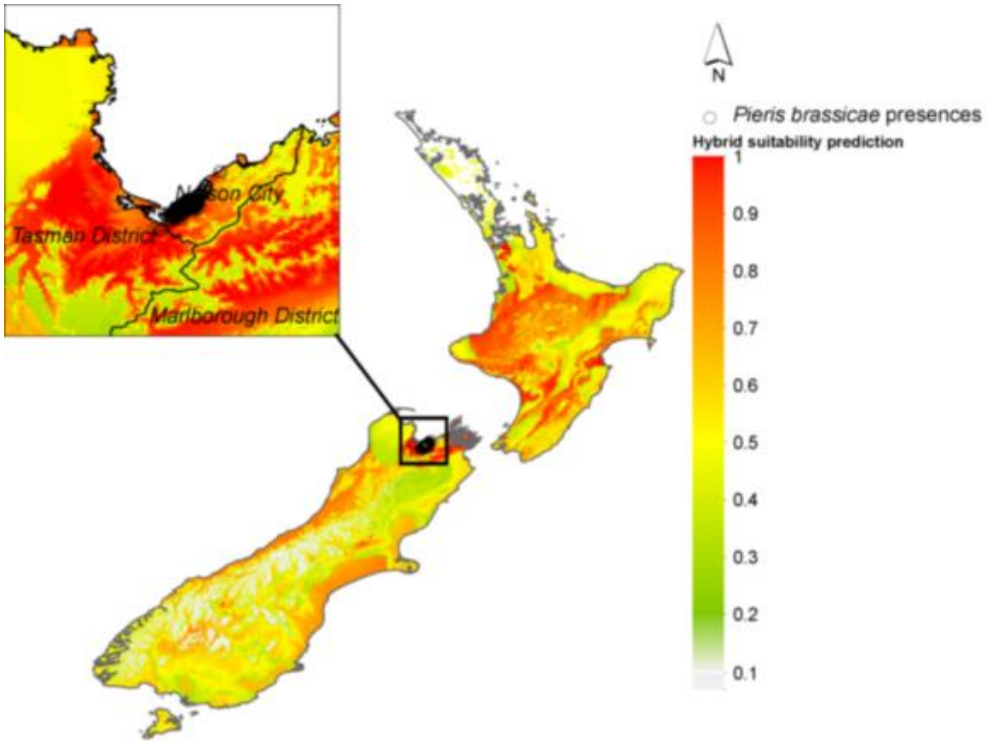


Figure 6.14: Hybrid suitability prediction for P. brassicae *in New Zealand. Inset maps show close up view of the area where* P. brassicae *is currently present. All coloured areas are suitable (predicted value >=0.5). Variation shows level of suitability.*

Although, the hybrid prediction represents only climatic suitability it can at least inform decision makers where climatic factors can aid to exacerbate *P. brassicae* dispersion additional to other factors. These factors include the availability of host plants especially

those of the *Cruciferae* family which were found to be the primary food preference of *P. brassicae* larvae (Ansari *et al*., 2012) and proximity to urban landmarks, as private gardens and untended communal green areas that have any of the alternative host species of *P. brassicae* can facilitate the establishment of *P. brassicae* populations even in the middle of an unsuitable landscape. Urban structures often condition the environment for invasive species to be favourable either by providing host plants or warmer climate (Lizée *et al*., 2011).

The decision to eradicate *P. brassicae* by the New Zealand department of conservation (DOC) was perhaps a very timely one as it would be more difficult to contain the spread of the species once it gets to the more suitable areas identified in Marlborough and Tasman districts that surround the Nelson city. Other highly suitable areas identified in the South Island include Buller, Grey, Westland, Selwyn and Waimakariri districts. Although, the suitable areas in these districts are further from where the species is established currently, the possibility of long distance dispersal through human aided transportation could put these areas at high risk. More suitable areas were identified in the North Island, some of the highest suitable areas were found in Franklin, Waikato, Waitomo, New Plymouth, Manawatu, Tararua, Taupo and Whakatane Districts. Even though, these areas are geographically well isolated from the *P. brassicae* population in Nelson, there might be a risk of introduction into North Island probably from human aided transport of diapausing pupae, as *P. brassicae* pupae are known to pupate on artificial substrates away from their host plants (Hagstrum & Subramanyam, 2010).

However, an in depth dispersal study that addresses multiple possibilities of *P. brassicae* movement by simulating different scenarios is needed to obtain more information on the future possible movements of the species. An additional study on the associated economic damage of *P. brassicae* is also necessary to determine if intervention is worth undertaking. On the other hand, if ecological implications are considered eradicating *P. brassicae* is most probably worth the cost as such aggressive invader is likely to be a great threat to the native butterfly community (Worner & Gevrey, 2006).

The potential distribution of *P. brassicae* produced in this study can be effectively used to produce a realistic landscape through a spatial analysis that considers factors like host

availability, species competition, physical barriers and landuse to further identify areas that cannot be occupied by the species for reasons other than environmental unsuitability.

6.4 Summary

6.4.1 Considerations about the SDMs used in this study

A major drawback in generalized simple mechanistic niche models is that seasonal variations of a given location cannot easily be incorporated (but see Monahan & Tingley, 2012 for their niche tracking study involving multiple seasonal niche projections using generalized niche models). In this study, a maximum mobility distance of 200 km was used to unify the geographic projections of two seasonal niches identified by the coldest and warmest months of the year respectively (Figure 6.3). Creating such connectivity between geographic areas that are seasonally suitable for invasive species allows identification of areas that might be under risk of invasion, even if they are not suitable all year round. A good example of such cases, is the migrant populations of *P. brassicae* that seasonally occupy the northern areas of the Scandinavian countries (Feltwell, 1982; Spieth & Cordes, 2012). Accounting for areas that could support such transient populations of an invasive species, because of their proximity to more suitable areas, have implication on the accuracy with which the potential distribution of the species is predicted (Sinclair *et al.*, 2010).

In generalized mechanistic models, varying limiting thresholds for the different life stages of the same species cannot easily be separately considered. When using simplified and generalized mechanistic niche models for an insect as employed in this study, it is important to investigate biological traits such as diapause or migration that could explain the persistence of the species in the environmental conditions that are close to the species lethal thermal limits. If valid assumptions based on an understanding of such traits can be made, using the upper and lower most lethal thresholds found in the literature allows the construction of a fundamental niche that is most likely to include all possible thermal combinations in which the species may persist (Monahan, 2009).

Regarding correlative models, the different scenarios showed that the models that are based on presences from the invaded ranges of *P. brassicae* predicted more suitable areas that are within the potential niche identified by the mechanistic model. Clearly, if the additional presence points from New Zealand were not included most of the areas in the North Island

of New Zealand would not have been predicted. Therefore, it is essential to report areas that are not predicted by correlative models as "no data" or "no information" instead of showing them as predicted absences which could be misleading for decision makers. Alternatively, if additional data like true absences, or a prediction from a mechanistic model is available, one can discriminate between under-predicted areas and actually unsuitable areas. For example, one can actually be confident about the northern parts of the Russian federation which was predicted unsuitable as it was outside the thermal thresholds of *P. brassicae,* more than other areas that were not predicted by the correlative model but were within the thermal threshold of *P. brassicae* (Figure 6.12).

One of the reasons the hybrid prediction was undertaken was to validate the combined prediction method introduced in Chapter 5. The New Zealand presence points of *P. brassica* have not been used in any of the models in Chapter 5. Yet, the prediction from the model that accounted for variation within presence data by modelling the aestivating and non-aestivating presences of *P. brassicae* separately, had identified most of the suitable areas predicted by the correlative model using the New Zealand points in this Chapter (Figure 5.12-B Vs. Figure 6.12-A). The central parts of the North Island, that were not predicted by the model in Chapter 5, show that the additional New Zealand points used in the pbPSNbt model in this chapter enabled identifying more suitable areas as discussed in the previous paragraph.

6.4.2 Predictability of a species

Prediction errors are usually associated with data inaccuracy, lack of model robustness, research design, validation errors or other numerous shortcomings of the ecological modelling processes (Elith *et al.*, 2002). While for the most part this is true, it is also important to consider the predictability of the species involved. It is very difficult to come up with a systematic standard index that measures predictability of a species. This difficulty is partially because there is not enough information on failed invasions to investigate if the target habitat has ever been invaded, not suitable enough or not occupied despite being suitable due to inadequate fitness of the invasive species. Even if such data exists, it is difficult to parameterize complex models that deal with such big data at a large scale. One way predictability of a species could be measured is by assessing the environmental distance

between where the species is currently present and its physiological limits in niche space (Miller & Stillman, 2012). It is easier to predict the potential distribution of a species that exists in the climate space closer to its physiological limits rather than a species that has large potential environmental range to fill before it hits its physiological limits. For example, the characterization the different fractions of the *P. brassicae* potential niche, showed that the current presences of the species are located far from its lethal thresholds in the thermal feature space. Such characterizations can be a general precursor to the invasion ability of the species. Characterizing uncertainty in potential species distribution predictions from biotic aspects will complement the on-going effort to account for sources of uncertainties in SDM predictions (Hartley *et al.*, 2006; Dormann *et al.*, 2008; Buisson *et al.*, 2010). Information on the predictability of a species will also help in the decision of whether to continue improving species distribution models or spend time in gathering more biotic information (Diamond *et al.*, 2012) to inform the models better on a case by case basis.

6.4.3 Setting rules to for hybrid correlative-mechanistic predictions

In previous studies, different methods have been used to combine mechanistic and correlative model outputs as hybrid models. The most used method is to incorporate the mechanistic model output along with other environmental covariates in the training datasets used for correlative models. (Buckley *et al.*, 2010; Elith *et al.*, 2010; Nabout *et al.*, 2012). The less explored method is using areas outside the potential niche as pseudo-absences for correlative models. This however, has the disadvantage of providing overly discriminated presence/pseudo-absence training datasets that leads to overfitting of models (Lobo *et al.*, 2010). If using such method is necessary it is better to take a portion of the pseudo-absences from outside the potential niche, and use additional pseudo-absence points that are in close proximity of presence points but are environmentally dissimilar from presence points. Such stratified sampling method was demonstrated by Elith *et al.* (2011) for sampling background data in MAXENT(presence-only model), however the principle behind the recommendation can be applied for selection of pseudo-absence points for presence-absence models.

In this study, a slightly different method was used to combine the correlative and mechanistic results. Instead of integrating the mechanistic output in the correlative

modelling process, the final predictions from the two models were combined using a simple rule.

Hybrid predictions that approximate the realized distribution are useful to manage species that are already in equilibrium with their realized niche. Such species usually have had time to uniformly occupy all areas that are suited to their requirement except in cases where they are excluded because of unfavourable biotic interactions or physical barriers (Gallien *et al.*, 2012). These interactions could be competition, host unavailability or parasitism. In such cases it is appropriate to take a model consensus approach to combine the outcomes of the mechanistic and correlative models, and extract only predictions that match. This method is more suited to conservation studies where higher specificity is required to identify best sites for relocation of endangered species.

On the contrary, a newly invading species is far from being at equilibrium with its available potential niche in the new habitat therefore it is important to identify the maximum possible boundaries of its potential distribution. This is important especially for a species like *P. brassicae* that is known for its ability to adapt to harsh environments. Such an invasive species can often occupy marginal habitats with environmental conditions that are outside their optimal thermal requirements (Beest *et al.*, 2013). Taking this into account the rule to hybridize the two outputs was set to maximize the probability of identifying all sites in which *P. brassicae* could establish. Even better hybrid models can be constructed by writing more complicated rules to consider positional and temporal uncertainties associated with the correlative and mechanistic models. Better hybrid predictions could also be achieved if the potential niche defined by the mechanistic niche model is based on an n-dimensional hyper volume constructed out of more environmental variables that directly affect the physiology of the species instead of only a two dimensional thermal feature space.

6.4.4 Concluding remarks

It is difficult to generalize mechanistic models, and that is the reason why more ways of improving correlative models will always be beneficial especially in cases where we do not have species specific information. However, if there is some physiological information on the target species, it is beneficial to at least construct a simple representation of the potential niche using generalized mechanistic niche models such as the ones in this study in order to

increase the confidence of species distribution predictions (Kearney *et al.*, 2010). This is essential when model outputs are to be directly used to plan surveillance and eradication programmes. Such projects often have significant costs and decision makers require a clear understanding of the varying risk of invasive species establishment so that they can prioritize their actions accordingly.

Generally, there are parts of a potential species distribution a correlative model might not predict even when it is optimized and improved because a portion of that environmental range is not represented in the training dataset (Diamond *et al.*, 2012). While correlative models can extrapolate outside the environmental range of the data they are trained on, the predictions cannot be taken at a face value due to lack of appropriate external validation data.

It is preferable to compare model results with mechanistic predictions to make sure all portions of the potential distribution are captured when possible. However, availability of biotic information on emergent invading species is not always readily available as species that are only important either economically or ecologically are studied more. In the event a new invasive species arrive at a new habitat it is less likely to find information in the literature especially if this species have not been found to be invasive in other areas. Therefore, It is important to continue to use improved and well specified correlative models where complete absence of biotic information about a species is encountered (Rowland *et al.*, 2011). It is also important to prepare a modelling framework that later incorporates physiological limits and requirements as such information becomes available to increase the level of confidence from correlative predictions.

Chapter 7

7. Landscape mapping for spatially explicit species dispersal models

7.1 Introduction

7.1.1 New Zealand invasion of Pieris brassicae

Pieris brassicae was introduced to the South Island of New Zealand in 2010 (MAF, 2012; DOC, 2013b). It was first reported by a home gardener in Nelson, and has since been detected in most parts of Nelson city. The butterfly also established outlying populations about 12 km south of central Nelson at Richmond, and also about 12 km north at Glenduan (Phillips *et al.*, 2013). An eradication campaign was initiated by the Department of Conservation (DOC) in September 2012. The eradication programme collects presence records through both passive and active surveillance (Phillips *et al.*, 2013).

Invasive species are generally unwanted and are rarely welcomed in any scenario. The introduction of *P. brassicae* in New Zealand, however, was especially alarming as the impacts of such destructive invasive species is even more pronounced on islands. This is because highly competitive invasive species could seriously damage island based ecosystems due to the fragility and uniqueness of the native fauna and flora found in Islands (Mooney & Cleland, 2001). The native cress species in New Zealand are potential hosts for *P. brassicae* and therefore are considered threatened. There are 79 native cress species in New Zealand and 57 of these are already at risk of extinction due to habitat loss and impacts from other pests (Phillips *et al.*, 2013). Even though DOC's focus is to protect native species, *P. brassicae* could also cause great economic damage as it feeds on commercially grown *Brassica* species. There are about 4,000 ha of brassica vegetable plantations in New Zealand and

250,000 ha of brassica forage crops, which is estimated to be worth 80 million New Zealand dollars (DOC, 2013a; Phillips *et al.*, 2013).

Following New Zealand's national Biosecurity Act, *P. brassicae* was deemed a threat soon after its discovery. A number of studies were quickly launched by the Plant and Food Crown Research Institute in collaboration with the Department of Conservation (DOC) and the Ministry of Primary Industries (MPI) (Kean & Phillips, 2013d- & references therein; Phillips *et al.*, 2013). These studies focussed on the biology and more specifically the phenology of *P. brassicae,* and provided a better understanding of how *P. brassicae* behaves in its new invaded range. These studies were also instrumental in designing surveillance and undertaking eradication programmes (Phillips *et al.*, 2013).

However, there are no studies of *P. brassicae* dispersal in New Zealand. Therefore, *P. brassicae* was chosen as a model species to investigate the effects of landscape mapping on results from spatially explicit species dispersal models. *P. brassicae* was also used as an example species in the previous two chapters, which produced global and New Zealand extent habitat suitability maps for the species. In contrast with the previous two chapters, the suitability layer constructed in this study has a much higher resolution and is composed of more than climate variables; this makes it suitable for local surveillance and eradication planning. I also characterised the dispersal dynamics of *P. brassicae* using data from the United Kingdom, which I chose both because of the availability of long time occurrence data, and due to the similar spatial extents of the United Kingdom and New Zealand. The resulting landscape map was used to test for the efficacy of the current eradication regime.

7.1.2 Invasive species-landscape interaction

Species distribution models based on climate variables are instrumental for understanding the habitats of invasive species at the global or regional scale. However, higher resolution and better detailed maps that adequately represent the landscape are necessary to inform invasive species surveillance and eradication operations at a local scale. One of the most obvious uses of such landscape data is characterising the heterogeneous surface over which spatially explicit dispersal models are used to simulate invaders movement (Pulliam *et al.*, 1992; Ruckelshaus *et al.*, 1997).

Species dispersal models have been used to design surveillance and eradication processes, evaluate outcomes of invasive species control and estimate damage caused by invasive species (Gilbert *et al.*, 2003; BenDor *et al.*, 2006; Lippitt *et al.*, 2008; Epanchin-Niell & Hastings, 2010; Økland *et al.*, 2010; Keith & Spring, 2013). Some dispersal models have also been used to generate various hypotheses regarding dispersal dynamics, species equilibrium and reconstructing past dispersal events (Keeling *et al.*, 2001; Acosta, 2002). There are different types of species dispersal models ranging from the early reaction-diffusion models that model uniform spread rate using partial differential equations as used by Fisher (1937) and Skellam (1951) to individual based models (IBMs) that simulate dispersal over either a homogeneous or heterogeneous landscape in a spatially explicit manner (UchmaDski & Grimm, 1996; Pitt *et al.*, 2009). In-depth reviews of types of species dispersal models including historical accounts and comparative studies are given by (Higgins *et al.*, 1996; Grimm, 1999; Hastings *et al.*, 2005; Pitt, 2008).

In this study, a spatially explicit modular dispersal model was chosen to investigate the use of selectively recoded landscape in dispersal predictions. To understand dispersal dynamics of a species, it is important to accurately estimate the species' dispersal capability (Okubo & Simon, 1989; Higgins & Richardson, 1999; Barney, 2006a). Additionally, it is also important to characterise the effect of the landscape on these capabilities. Spatially explicit models designed over heterogeneous landscapes have the advantage of accounting for the effect of the landscape on species dispersal dynamics (Higgins *et al.*, 1996). This is due to the effect of landscape characteristics on the mobility and survival of invading species (Ewers & Didham, 2006). The appropriate specification of spatio-temporal aspects of species' dispersal characteristics, improves the characterization of dispersal patterns (patchiness) (Cavanaugh *et al.*, 2014) and the accuracy of dispersal rate estimates (Hastings *et al.*, 2005).

It is sometimes possible to model patchy dispersal even without a landscape input. For example, Morozov *et al.* (2008) modelled a patchy distribution in a non-spatial model of a multi species dispersal system. However, their system was a multi-species scenario where interactions between varying inter-species dispersal parameters can facilitate patchy dispersal (Neubert *et al.*, 1995). Thus, it is possible to have patchy distributions regardless of

environmental inputs, though landscape will still have additional influences on dispersal dynamics (Morozov *et al.*, 2008).

Apart from realistically representing dispersal patterns, spatially explicit models reveal the mechanics of dispersal dynamics better than non-spatial models. This means in addition to better definition of a dispersal pattern at the end of any given discrete time, *t*, the effect of the landscape on the rate, direction and success of future dispersal could also be easily represented. For example, suitable patches in the middle of unsuitable areas can act as continuous sources of propagules, replenishing unsuitable sites where the species would otherwise be extinct (Pitt *et al.*, 2009).

For studies that aim to inform invasive species management plans, the most important factors to consider for dispersal model choice are the landscape and the species to be modelled. If the spread of an invasive species is modelled for monoculture plantations or crops or greenhouse experiments the mathematical models which assume a homogeneous environment, but can incorporate complex assumptions about the effect of competition and other factors on dispersal dynamics, might be preferable (Pitt *et al.*, 2009). On the other hand if the area for which the study is being designed is heterogeneous, and especially if it is interdispersed with indigenous fauna and flora then a spatially explicit modelling approach that considers heterogeneity is appropriate to prioritize intervention sites, and assess the possible effect of an invasive species as well as the eradication effort on local fauna and flora. Species specific information also affects the type of landscape to consider. For example, dispersal parameters of species that actively seek their suitable environment over a long distance might be more influenced by the configuration of the landscape, than those that strictly depend on another agent for their long distance dispersal (Schellhorn *et al.*, 2014). While modelling such a species it is important to understand the configuration of the landscape to understand its possible effect on the spread rate of the species. One such species is the large white butterfly (*Pieris brassicae).* The female adult is known to fly over 200-400 km after emergence (Spieth & Cordes, 2012), actively seeking *Brassica spp* or other alternate host plants (Chew & Renwick, 1995).

7.2 Methods

"That there is no single correct scale or level at which to describe a system does not mean that all scales serve equally well or that there are not scaling laws."

Simon A. Levin in his MacArthur address, 1992

The last objective of this thesis, to investigate recoding heterogeneous landscapes with a varying degree of composition across space to improve dispersal rate and pattern predictions, is investigated based on methods provided in the following sections.

7.2.1 Study area

Determining the landscape extent and resolution for insect dispersal studies is a subject of ongoing research (Turner *et al.*, 2001; Skelsey *et al.*, 2013). There is no spatial extent and/or resolution that are known to universally optimize dispersion for all species (Levin, 1992; Chave & Levin, 2004). It is important to consider specific species biotic parameters especially the local spread rate, dispersal pathways, and the type of the landscape when setting spatial resolution for a dispersal model (Higgins *et al.*, 1996). Spatial extent is usually defined by the research aims, the available computation power, spatial resolution, and abiotic information. While the factors given above relate to the scope of this study, the choice of spatial scale and extent spans many more spatio-temporal factors than discussed here (Saura & Martinez-Millan, 2001).

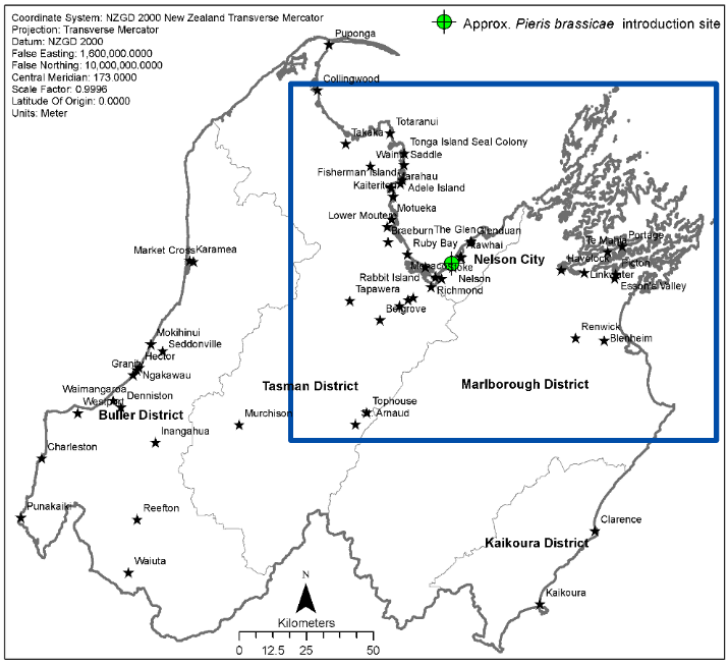


Figure 7.1: Study area of the P. brassicae *dispersal modelling study To avoid the edge effect in processing data all available datasets that were selected for the dispersal modelling were first clipped to cover the five administrative districts in the South Island that were either in contact with- or nearby the locations invaded by* P. brassicae. *These districts were Buller, Tasman, Nelson City, Marlbourgh and Kaikoura. This greater extent area of interest was used to process the various inputs for the dispersal model before clipping all the necessary input data to the final extent of the study area which was 12, 466 sq. km bounded by the blue box in Figure 7.1. The approximate location of the initial introduction point of* P. brassicae *is located near the middle of the study area.*

7.2.2 Biological aspects

P. brassicae is a known migratory species that is reported to be capable of flying up to 400 km (Spieth & Cordes, 2012). Although wind could assist insects in their long flights (Pasek, 1988), *P. brassicae* has been observed flying upwind (Williams, 1958). This suggests that wind is not necessarily the cause for *P. brassicae* long distance dispersal, and that it may simply take advantage of wind when its direction is suitable (Feltwell, 1982). Adult *P. brassicae* tend to fly in one direction after emerging and the fact that there are usually no return flights to their starting point increases the possibility of travelling long distances. However, it is reported that the offspring usually find their way back to where their parents originated (Spieth & Cordes, 2012) allowing panmictic populations across their range and providing reinforcement in areas *P. brassicae* populations are already present (Hansen & Merivee, 1971). A female adult *P. brassicae* will only alight and rest either to wait out a strong wind blowing against her direction (Williams, 1958), to feed or to lay eggs if she detects a suitable host under her flight path (Feltwell, 1982).

P. brassicae is reported to have a proportional (1:1) sex ratio during migratory long flights (Feltwell, 1982), which can reduce the Allee effect on migrating *P. brassicae* groups. This has important implications for characterising the long distance dispersal of *P. brassicae,* as reduced survival consideration to account for an Alee effect can lead to under-estimation of dispersal capability for this species.

It has been reported that *P. brassicae* detects hosts using a combination of visual and biochemical mechanisms. A *P. brassicae* in flight, whether in a group or alone, will first detect a host using colour cues (Rothschild & Schoonhoven, 1977; Jolivet, 1992). Upon nearing the vegetation, they use various hormonal and biochemical mechanisms (Debarma & Firake, 2013) to check the suitability of the plant for oviposition. As *P. brassicae* does not tend its offspring, the female must optimise the survival of her offspring by selecting appropriate host plants.

Even though *P. brassicae* can disperse long distances on its own, humans are responsible for transferring *P. brassicae* across oceans or other large areas of unsuitable habitat. The life stages that are mostly associated with human assisted dispersal are eggs and pupae. This is because *P. brassicae* pupates on artificial substrates like walls, cargo containers, fence and

posts. rather than on host plants. This creates the possibility that pupae are accidentally transported through industrial, commercial or public transportation networks. There was an unconfirmed report where 400 pupae were discovered off containers shipped from Spain at the Colorado port (NAPPO, 2002). *P. brassicae* pupae have also been intercepted at the New Zealand border by the MPI, one on a caravan and another on a car (Craig Phillips, Pers. Comm. March 14, 2014). It is suspected that *P. brassicae* could have been introduced to Nelson with shipping containers off-loaded at the Nelson port. While transportation of pupae seems to be the major pathway for cross-border introduction, eggs can also be moved with host plants such as commercial brassicae species; this is probably more important for national rather than international dispersal.

7.2.3 Spatially explicit dispersal model- MDiG

In this study the Modular Dispersal in GIS[32] (MDiG) model developed by Pitt (2008) was used. MDiG is an open source program that is developed within the framework of another open source GIS program GRASS (Geographic Resources Analysis Support System) (Neteler *et al.*, 2012). This has the advantage that users can adapt the program to their needs, and that it is fully integrated with a powerful GIS system makes it easy to export any outputs in the form of GIS maps or database for subsequent spatial analysis or reporting.

MDiG was developed to handle different kinds of dispersal mechanisms in a step by step manner. This enables accounting for the most important dispersal parameters like population growth, long distance dispersal and species-landscape interactions through its various modules. This design actually mimics the natural way of species dispersal, as individuals spread by one mode of dispersal at a time, but the same individual can use more than one mode of dispersal in their life time (Mader *et al.*, 1990; Bilton *et al.*, 2001; Pitt, 2008). While MDiG is a spatially explicit individual based population model that is capable both of characterising different life stages with different parameters and of estimating abundance-related dispersal, in this study the presence/absence option was used. This kept the model simple so the effect of landscape on dispersal could be more easily studied.

Here, only the modules used in the present study are described. However, MDiG is a much more powerful model and has other functionalities beyond the modules described here. A

[32] Geographic information systems

detailed description can be found on the official site for the program (http://fruitionnz.com/mdig/) and Pitt's (2008) doctoral thesis.

7.2.3.1 Local spread through the neighbourhood module

The neighbourhood module given as "`r.mdig.neighbour`" is designed to represent local spread from contagious cells. The two parameters that affect local dispersal dynamics are shape and range (Pitt, 2008). Shape is the pre-determined direction and order of cells an occupied cell can infest in its locality at any one time step. The Von Neumann and the Moore neighbourhood are the shapes used in most neighbourhood analysis computations, including cellular automata and other raster based focal analysis (Figure 7.2).

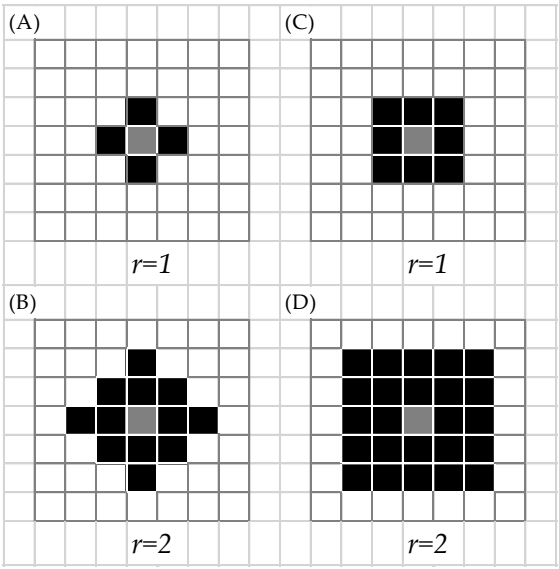


Figure 7.2: Local dispersal neighbourhoods, (A) Von Neumann shape with range = 1 (B) Von Neumann shape with range = 2 (C) Moore shape with range = 1 (D) Moore shape with range = 2

7.2.3.2 Kernel dispersal

The kernel module, given as "`r.mdig.kernel`" addresses long distance dispersal. It stochastically generates long distance events, hence there is no explicit method to integrate different dispersal pathways separately. For example, a species could disperse by wind, through human transport or by active flight or by walking. Although, representing such pathways explicitly would generate more realistic long distance dispersals, it is usually very difficult to find such detailed information on dispersal pathways for understudied species. If such information was available, then dispersal kernels characterised by mixed probability distributions could be used to generate such events (Gilbert *et al.*, 2004). The MDiG kernel module samples a Poisson distribution to determine the number of dispersal events that are

generated from an occupied site. Once the number of events is determined, a user defined probability distribution is sampled to determine the distances of the events generated from the occupied site. The options of probability distributions included are, Cauchy, exponential, logarithmic and a general kernel that users can characterise by defining the shape and scale parameter (Pitt, 2008). Finally, a uniform distribution in the range of $[0,2\pi]$ is sampled to determine the direction of each generated long distance events (Pitt *et al.*, 2009).

7.2.3.3 Survival module

The survival module given as "`r.mdig.survival`" in the program allows a species-landscape interaction to be incorporated. The user specifies a habitat suitability index in the form of survival probability map ranging from 0-100. The individuals modelled through the local and kernel modules are passed through the survival module to determine the ones that survive based on the underlying suitability value. It is also possible to provide a single survival value if the landscape is homogeneous such as with a monoculture glasshouse (Pitt, 2008).

7.2.4 Parameterizing P. brassicae *dispersal*

P. brassicae exhibits both local and long distance dispersal (Feltwell, 1982), which were parameterized as follows.

7.2.4.1 Local dispersal

The neighbourhood module of MDiG was used to define the local dispersal of *P. brassicae*. Unlike long distance dispersal, there is no evidence that *P. brassicae* exhibits directional bias during local dispersal (Davies & Gilbert, 1985). Thus, the Von Neumann shape (Figure 7.2) with range = 1 was chosen to represent the neighbourhood for spread within one time step because it has a more or less uniform direction. Although, the Moore shape has a more uniform direction, it can lead to over-estimation of local dispersal because it has more number of cells that expand within one time step.

I chose a cell resolution of 100 m to approximate local movements of larvae and adults. A number of estimates including 7 m, 88 m and 350 m were reported as the distance *P. brassicae* larvae cover while looking for pupation sites, as reviewed by Feltwell (1982), with the latter reported as unusual. I took the median value of the distances reported for larval

movement above, and rounded it to the nearest hundred meters to make area calculations straightforward. I refrained from using spread rates reported at large scales, which included long distance dispersals, as these are dealt with by the kernel module. Also, estimating local spread rates using long-distance aided dispersal estimates is misleading as it does not give reliable local scale information (Neubert & Parker, 2004). The initial dispersal site was set in a cell close to Nelson port which is suspected to be the site of *P. brassicae* introduction.

7.2.4.2 Source data for long distance dispersal parameters

The dispersal history of *P. brassicae* in the United Kingdom has been characterised, and the resulting estimates were used to parameterize the long distance dispersal module in this study. *P. brassicae* has a Palearctic distribution. The reasons why United Kingdom is chosen to parameterize long distance dispersal are:

1) *Surface area: the United Kingdom is an island territory with a similar land surface area to New Zealand. The kernel module does not address each long distance dispersal pathway separately, but represents the overall dispersal pattern in terms of distance, frequency and direction. Therefore, obtaining dispersal parameters from an area similar to the study area improves predictions. For example, a factor that affects dispersal simulations is spatial extent; with limited spatial extent, some long distance dispersal events are lost because they fall outside the study area. Some of these lost events could have been sources of dispersals back into the study area, and this introduces error in terms of both the final predictions as well as the dispersal dynamics (Pitt, 2008). While error is inevitable as one can only model a limited area at a given resolution provided the time, data and computation resource limitations, it is possible to reduce it (Baker et al., 1995). Here, I reduced error arising from spatial extent by using a source area (United Kingdom) for parameter estimation similar to the study area (New Zealand).*
2) *Data availability: It is usually difficult to find geographically accurate species occurrence data, and it is even more difficult to find species occurrence data complete with temporal information. There are over 103,059 geographically referenced high resolution global occurrence points for* P. brassicae *on the GBIF database. However, only a few contain dates of introduction of* P. brassicae *at the locality it was recorded.*

Some *P. brassicae* data from the United Kingdom was well referenced in terms of temporal records. Additional information on pre-1960 distribution of *P. brassicae* was obtained from Feltwell (1982) and Heath (1970).

7.2.4.3 Distance estimation for long distance dispersal events

A map of *P. brassicae* distribution in the United Kingdom (Feltwell, 1982) was scanned and rectified using the ArcGIS® software according to the UTM Zone 30 projection information

given in the legend. The provisional version of the original atlas Heath (1970) was downloaded from the Open Research Archive of Natural Environment Research Council website because it had clear map ticks that could be used for the geo-rectifying process. The atlas described the sampling scheme used to collect the distribution data and indicated that each occurrence point represented a 10 km sq. grid on the map. Accordingly, a cell size of 10 km was used as the standard data processing resolution for dispersal parameter estimation. The *P. brassicae* occurrence points were then digitized from the rectified image for the pre-1960 (n=57) and post-1960 (n= 294) time periods given in the legend.

Both the pre-1960 and the post-1960 occurrence datasets, and the United Kingdom boundary traced from the rectified scanned image, were projected to the British National Grid coordinate system. National grids give the most accurate planar distances locally.

The boundary traced from the rectified image was overlaid with a standard national boundary dataset obtained from the United Kingdom Ordinance Survey website. The total area of the spatial difference between the boundary of the rectified image and the standard boundary data was estimated after the overlay operation by extracting non-conforming areas. The area of the difference-data was divided by the area of the standard UK boundary to obtain the approximate uncertainty per km sq. from the rectification process.

From the GBIF database, 56,913 *P. brassicae* occurrence points were found for the United Kingdom. Most of the points conformed to the sampling grid of 10 km sq. used in Heath (1970). This suggests continuity of the survey as well as overlap between Heath (1970) and the GBIF database. However, a high density of occurrence points that were tightly localized in the counties of Cumbria, Shropshire, Bath and North East Somerset, North Somerset, South Gloucestershire and City of Bristol conformed to high resolution (1 km sq. grid) sampling units. These highly localized urban surveys inflated the number of occurrences (Appendix 7.2). To remove bias due to intensive surveying in for these areas, the GBIF data were resampled using 10 km sq. intervals, and were projected into the British national grid coordinate system. The resampled GBIF *P. brassicae* occurrence data (n=505) along with the points digitized from Heath (1970) (n = 351) were used to estimate dispersal parameters for *P. brassicae*.

A total of 856 points were used for parameter estimation. Information on dates of records of *P. brassicae* occurrences from Feltwell (1982) and GBIF database was used to characterise the 856 points with a year of introduction into their respective localities. For example, for the pre-1960 dataset the occurrence points were characterised by the year of their first sighting to give the data further temporal resolution.

Two distance measurement methods are usually used to estimate the distances travelled during long distance (jump) dispersal of invading species; either a nearest neighbour method, or a distance from the origin method (Robinet *et al.*, 2009). The distance-from-the-origin method, measures distances from each occurrence point to the site of estimated origin of introduction of the species into the study area. While this works best for small scale studies, it over-estimates the dispersal distances in cases where the dispersal behaviour of the species is expected to be stratified (local plus long distance dispersal), long distance dispersers could possibly arise both from the front as well as the core of the invading population (Nathan & Muller-Landau, 2000). The dispersal distance for *P. brassicae* was parameterized by producing the nearest neighbour distance vector between occurrence points for each period to account for the stratified nature of *P. brassicae* dispersal. A random uniform noise within a range of [-146 m, 146 m] was applied to the nearest neighbour distance vector extracted from the occurrence points with many replications (n=1000) to account for uncertainty from the digitized points (Pitt *et al.*, 2011).

P. brassicae dispersal was characterized by a stratified dispersal where a population invades an area and locally spreads. As the density of the population within an area increases *P. brasscae* females fly away to find host plants that are not already burdened by eggs or larvae; this implies density dependent long distance dispersal (Rothschild & Schoonhoven, 1977; Debarma & Firake, 2013). Such dispersal is characterized by a high number of small distance dispersals followed by rare long distance dispersals as adults look for additional suitable hosts once pre-existing sites are over-populated. The Cauchy distribution was chosen to fit the long distance dispersal data due to the fat-tailed characteristics of the distribution that allows for rare long distance events (Kot *et al.*, 1996; Higgins & Richardson, 1999; Cain *et al.*, 2000). The Cauchy probability density function is given below.

$$f(x|x_o,\gamma) = \frac{1}{\pi}\left[\frac{\gamma}{(x-x_o)^2+\gamma^2}\right] \quad Eq.\,7.1$$

Where x_o is the location parameter and γ is the scale parameter of the Cauchy distribution.

The location (x_o) and scale (γ) parameters were estimated by fitting the noised distance data (n=1000) to the Cauchy distribution using the maximum likelihood estimator with the trust-region-reflective optimization algorithm (Conn *et al.*, 2000) using MatLab® software. The mean and standard deviation of the parameter estimates over the 1000 replicates was used to assess the stability of the estimated Cauchy location and scale parameters.

The New Zealand distance data were also analysed to compare the parameters from the two locations. However, *P. brassicae* was introduced to New Zealand only in 2010, so New Zealand dispersal data cannot fully represent the species dispersal dynamics. Additionally, New Zealand data will be biased towards short distances due to intensive surveillance and eradication conducted in 2012 and 2013. The New Zealand data, however was used to validate simulated eradication on dispersal results for New Zealand.

7.2.4.4 Frequency estimation

The frequency of the dispersal events was estimated from the United Kingdom temporal occurrence data. The historical *P. brassicae* presence data was first classed into periods. The ratio of the number of new *P. brassicae* sites to the number of existing sites was calculated for each period. The resulting vectors of ratios that reflect the minimum number of dispersal events that needed to be generated from each cell to achieve the number of occupied cells in the next time period were then weighted by the number of sites for each period as described by Pitt *et al.* (2011) (see Eq. 7.2).

$$\mathrm{R} = \frac{\sum_{t=t_{0+1}}^{T} N_t - N_{t-1}}{\sum_{t=t_{0+1}}^{T} N_{t-1}} \quad Eq.\,7.2$$

Where R is the vector of average weighted ratios calculated by dividing the number of newly invaded cells to the number of existing cells, t_0 is the first year with occurrence data, T is the last year, and N_t is the number of cells that are occupied within time t (Pitt et al., 2011).

The vector R which was made up of the estimations of the number of dispersal events per cell per year from the historical occurrence data was obtained from Eq. 7.2 and was fitted to

the Poisson distribution where the expected mean frequency of the distribution (λ) was estimated. The Poisson discrete probability function is given in Eq. 7.3.

$$f(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \quad Eq.7.3$$

Where k is a vector of discrete integers [0, 1, 2, …] that correspond to the a discrete random variable X whose probability mass function is given by Eq. 7.3., and λ is the expected mean of the distribution with $\lambda > 0$ condition. The λ in this case is estimated by fitting the vector R from Eq.7.2 to a Poisson distribution.

The obtained expected mean (λ) was used to parameterize the Poisson distribution used by the MDiG kernel module to determine the number of events generated from each occupied cell. The number of dispersal events from each existing cell per year was also calculated for New Zealand to compare with the United Kingdom data.

7.2.5 Building the survival layer

A survival layer was used in the dispersal procedure to determine whether dispersing *P. brassicae* will survive in newly occupied cells. Two survival layers were developed to investigate the effect of landscape on the initial condition of invasive species dispersal. The first survival layer (Surv_1) was prepared using four data sources: climate suitability, degree days, land cover and high resolution remotely sensed data. The second survival layer (Surv_2) included all components used in Surv_1, except the high resolution remotely sensed data. Brief explanations of these datasets are given below.

7.2.5.1 Climate suitability

The hybrid climate suitability layer (Figure 7.3-A) produced in Chapter 6 was used as a component to construct a survival layer for the MDiG dispersal model. This layer was produced by combining correlative and mechanistic model results to obtain climate suitability prediction for *P. brassicae*. The data were projected to New Zealand Geographic Datum 2000 coordinate system, which was used throughout the study in this Chapter. The hybrid climate suitability layer had a resolution of 30′ (converts to ~ 0.8 km at latitudes 40° - 50°). Therefore, it was resampled at 100 m which was the raster resolution set for the dispersal model.

7.2.5.2 Growing degree days

A 30 year daily minimum and maximum temperature data obtained from NIWA (2005) was used to prepare the growing degree days map for *P. brassicae* in New Zealand. The base development temperature for the various life stages of *P. brassicae* were obtained from Kean and Phillips (2013c), which has detailed records of the individual life stages and total degree day requirement by *P. brassicae* to complete a generation in New Zealand. Their study developed a phenological model, partly parameterized by field data obtained observations of *P. brassicae* in New Zealand. These results based on local information were the appropriate input for a dispersal study aimed at the same locality. Base temperatures (lower threshold for development) of 8.2 °C, 17.9 °C, 10 °C and 11 °C were reported for the development of *P. brassicae* eggs, larvae, pupae and adults respectively. Because a presence/absence scheme rather than a detailed life stage dispersal was used in this study, it was necessary to choose one base temperature. The pupal development threshold was selected as a base temperature because: 1) pupa has the highest temperature development threshold of all the life stages. Thus, it's impossible for GWB to complete its lifecycle unless temperatures exceed this threshold (Craig Philips, Pers. Comm. March 14, 2014), 2) it is a stationary stage which is important as mobile life stages can move to more suitable locations, 3) it is close to the average baseline temperature for all the life stages compared to the other stationary life stage which is the egg.

The growing degree day (GDD) surface was produced by interpolating grid data from the accumulated GDD point dataset calculated using the base temperature 10 °C based on the 30 years daily minimum and maximum temperature data for New Zealand. There are a number of methods to calculate degree days and each method depends on the kind of temperature data available (Worner, 1988, 1992). The Barlow (Barlow & Dixon, 1980) or 4-step method was used to calculate the daily degree day value in this study as it was reported to be one of the methods that gave the least error when validated with real data (Kean, 2013). The daily average for the Barlow calculation was made by taking the average of daily maximum and minimum temperatures as true daily mean temperature data was not available. The degree day data calculated for 509 points was interpolated into a raster surface using spline interpolation (Schoenberg, 1971) (Figure 7.3-C).

$$dd = \frac{([\,T_{max} - b] + [\,T_{min} - b] + 2[\,m - b])}{4} \quad Eq.7.4$$

Where T_{max} is the daily maximum temperature, T_{min} is the daily minimum temperature, b is the base temperature and m is the daily mean temperature given by (T_{max}-T_{min}/2) or given by a true daily mean calculated from hourly or higher temporal resolution temperature measurements.

Table 7.1 Land cover re-classification according suitability for *P. brassicae*

Land cover name	Code	description
Built-up Area*	1	Very high suitability
Orchard and Other Perennial Crops	1	Very high suitability
Short-rotation Cropland	1	Very high suitability
Urban Parkland/ Open Space	1	Very high suitability
Alpine Grass-/Herbfield	2	high suitability
Depleted Tussock Grassland	2	high suitability
High Producing Exotic Grassland	2	high suitability
Low Producing Grassland	2	high suitability
Tall Tussock Grassland	2	high suitability
Broadleaved Indigenous Hardwoods	3	moderate suitability
Deciduous Hardwoods	3	moderate suitability
Manuka and or Kanuka	3	moderate suitability
Matagouri	3	moderate suitability
Gorse and Broom	4	low suitability
Herbaceous Freshwater Vegetation	4	low suitability
Herbaceous Saline Vegetation	4	low suitability
Indigenous Forest	4	low suitability
Major Shelterbelts	4	low suitability
Mixed Exotic Shrubland	4	low suitability
Sub Alpine Shrubland	4	low suitability
Vineyard	4	low suitability
Afforestation (imaged post LCDB 1)	5	very low suitability
Afforestation (not imaged)	5	very low suitability
Fernland	5	very low suitability
Flaxland	5	very low suitability
Forest Harvested	5	very low suitability
Grey Scrub	5	very low suitability
Other Exotic Forest	5	very low suitability
Pine Forest - Closed Canopy	5	very low suitability
Pine Forest - Open Canopy	5	very low suitability
Alpine Gravel and Rock	6	not suitable
Coastal Sand and Gravel	6	not suitable
Dump	6	not suitable
Estuarine Open Water	6	not suitable
Lake and Pond	6	not suitable
Landslide	6	not suitable
Permanent Snow and Ice	6	not suitable
River	6	not suitable
River and Lakeshore Gravel and Rock	6	not suitable
Surface Mine	6	not suitable
Transport Infrastructure	6	not suitable

** "Built-up Area" in the LCDB2 land cover dataset generally refers to urban areas or settlements and as it is a large scale data, it does generalize inner city gardens or small green spaces.*

7.2.5.3 Land cover/land use maps

The New Zealand Land Cover Dataset, LCBD2 (MFE, 2004) was accessed from the data portal Koordinates.com to extract land cover data for the study area. The land cover data was produced by Landcare Research based on SPOT imagery (resolution 15 m) and the pan-sharpened Landsat 7 ETM+ imagery (resolution 15 m). The dataset has 43 types of land covers. These were grouped and re-classed into six classes according to their suitability for *P. brassicae* (Table 7.1). The re-classed ESRI® polygon dataset was then converted to raster using 100 m resolution (Figure 7.3-B).

7.3.5.4 Selective enhancement of landscape detail using remotely sensed data

Remote sensing data is used for a wide array of agricultural and ecological research (Kerr & Ostrovsky, 2003). Satellite images are one form of remotely sensed data that are becoming increasingly available and affordable. The types of information extracted from satellite images differ based on the spectral and spatial range and resolutions of the image (Schaepman *et al.*, 2009). The objective of using remotely sensed data in this study was to characterise specific attributes of the urban areas within the study area including Nelson city where *P. brassicae* was first reported in New Zealand.

Here I carried out selective recoding of the landscape using image derived data out of the necessity to define a better initial dispersal condition for *P. brassicae* in Nelson. The need for such landscape arose because the land cover dataset available has classified all urban areas as built-up area (Table 7.1) which combined the highly suitable home gardens, public parks and untended green spaces with the houses, roads, buildings and other urban structures. And because *P. brassicae* is spreading in Nelson by breeding in home gardens that are available in residential blocks, a homogeneous landscape (Figure 7.4-A) within urban areas will over-estimate its dispersal. The remotely sensed data was used to characterise the geographical detail in urban areas in the study area. A single layer labelled "man-made structures" (Figure 7.4-C) was generated from the satellite image to update the homogeneous "built-up" (Figure 7.4-A) class in the land cover data.

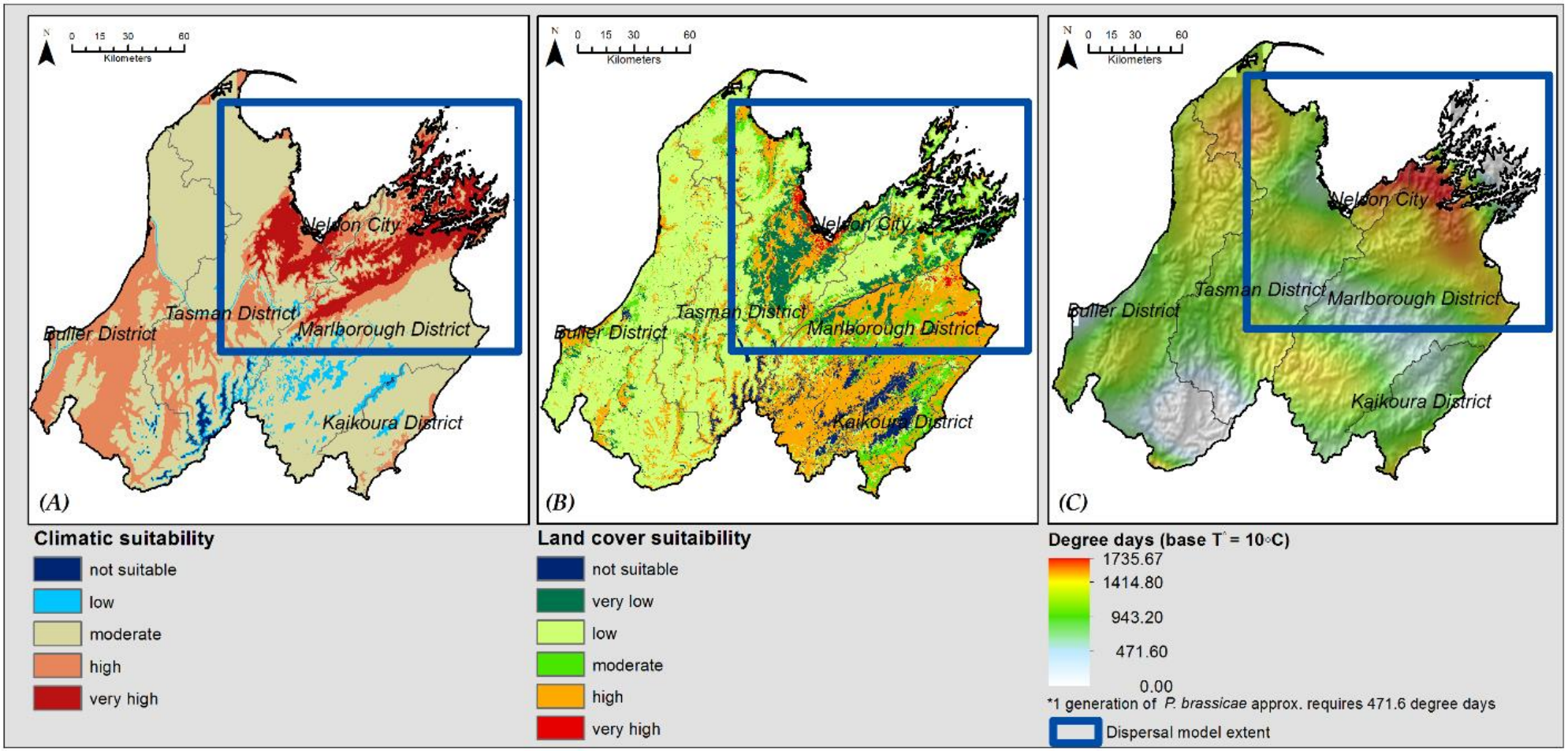


Figure 7.3: Suitability maps used to build the survival layer:(A) Hybrid climate suitability, (B) Land cover suitability and (C) Accumulated growing degree days

The chosen satellite image for this purpose was the SPOT[33] Maps® 2.5 m resolution satellite imagery. The image had a pseudo-colour band and three layers in the visible band, from which vegetation and man-made structures could easily be identified. The 2.5 m spatial resolution also enabled to identify patches of unsuitable land within urban areas with high precision.

Unsupervised image classification was first undertaken to roughly identify spectral reflectance associated with man-made structures like houses, buildings, roads and dams. All reflectance that identified such features were then merged to form one class for man-made structures (Figure 7.4-C). Ground truth data points (n=200) were collected from various GIS datasets including New Zealand road centre lines, New Zealand Rivers and land cover data to validate the unsupervised SPOT image classification.

The image classification result is briefly discussed here and is not presented in the results section. The accuracy of the classification was assessed using a confusion matrix and was 92.5%. This high level of accuracy was expected as the features being classified were spectrally very distinct and the focus was on getting precision in only one class (man-made structures) which has features very distinctly separated from the other land covers like crops, water and vegetation. The few miss-classifications associated with the commission error were heavily logged or deforested patches that had similar signatures to open roads. The omission errors occurred where side vegetation obscured the view of roads on the satellite image, were classified as non- man-made class.

7.2.5.5 Survival layer combining scheme

The method to combine these different components of the survival layer was designed so that the most limiting factor to *P. brassicae* dispersal had the highest weight. The climate suitability dataset was downscaled from its original 0.8 km resolution to 100 m which means change in value is not expected at least for eight cells (pixels). Therefore, the least weight was given for the climate suitability dataset. In Chapter 6, it was shown that temperature

[33] Satellite Pour l'Observation de la Terre

throughout New Zealand are generally suitable for *P. brassicae*, though some areas are more suitable than others. Hence, the growing degree day data was given medium weighting.

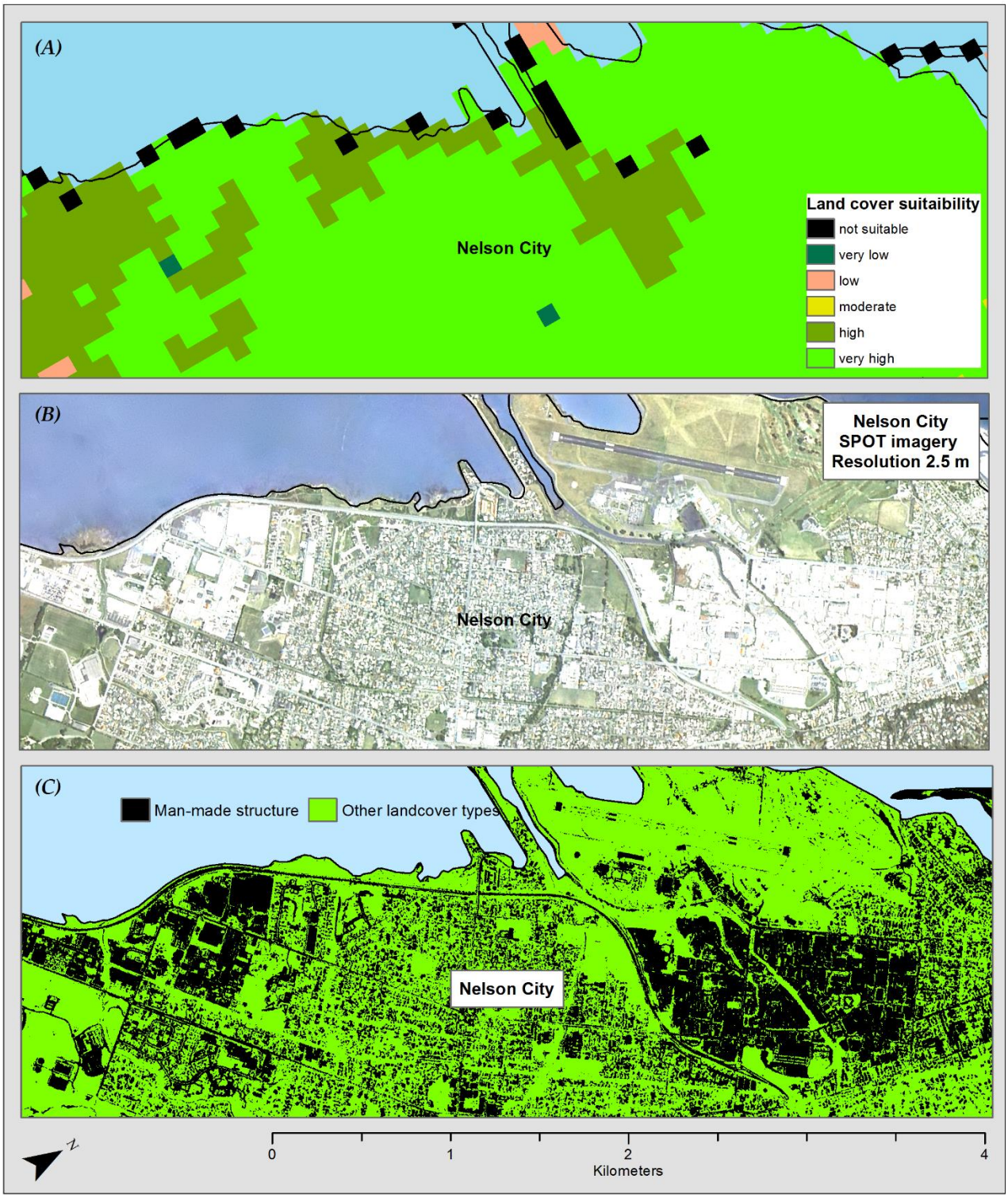


Figure 7.4: Comparison of geographic detail for urban areas between the land cover dataset (A) the high resolution SPOT Map® imagery (B) and the classified man-made structure layer (C).

The land cover data was derived from a 15 m resolution image and showed the highest longitudinal variation of all the layers, so was given the highest weight. The land cover data was used as the base on which extra values from the above two suitability layers was either added or subtracted depending on their assigned suitability value (Table 7.2). The final survival layer was then clipped to the study area extent set for the dispersal model.

Two survival layers were produced. The first one included the climate suitability, degree-day suitability, land cover and the man-made structure extracted from the satellite image and was labelled as Surv_1. The second survival layer included all layers except in Surv_1 except the man-made structure layer and was labelled Surv_2.

Table 7.2: scheme used to combine different survival layer components

Land cover		Climate suitability (%)		Accumulated degree day		Man-made structure (Boolean)	
code	Survival probability	Value	Contribution *	Value	Contribution	Value	Contribution
6	0	< 10	- 10	< 471	- 10	If 1	Set survival to zero
5	10	10 - 20	- 8	471 - 942	+ 4	If 0	No input
4	30	20 - 30	- 6	942 - 1413	+ 6	-	-
3	50	30 - 40	- 4	1413 - 1884	+ 8	-	-
2	80	40 - 50	- 2	> 1884	+10	-	-
1	90	50 - 60	+ 2	-	-	-	-
-	-	60 - 70	+ 4	-	-	-	-
-	-	70 -80	+ 6	-	-	-	-
-	-	80 – 90	+ 8	-	-	-	-
-	-	> 90	+ 10	-	-	-	-

**Contribution here refers to the percentage that gets added or subtracted to/from the survival layer depending on the level of suitability of each component. Land cover was taken as the base survival layer on which the effect of the other components is added and is given in the first column.*

In case of the first survival layer (surv_1) the final combined survival layer was recoded with the man-made structures data identified from the SPOT image classification where all areas that were overlaid by the man-made structure data were set to a survival of zero probability (Figure 7.5-A). For the second survival layer (surv_2) the dataset that contained the combination of the three different suitability datasets was used without the SPOT image data input (Figure 7.5-B).

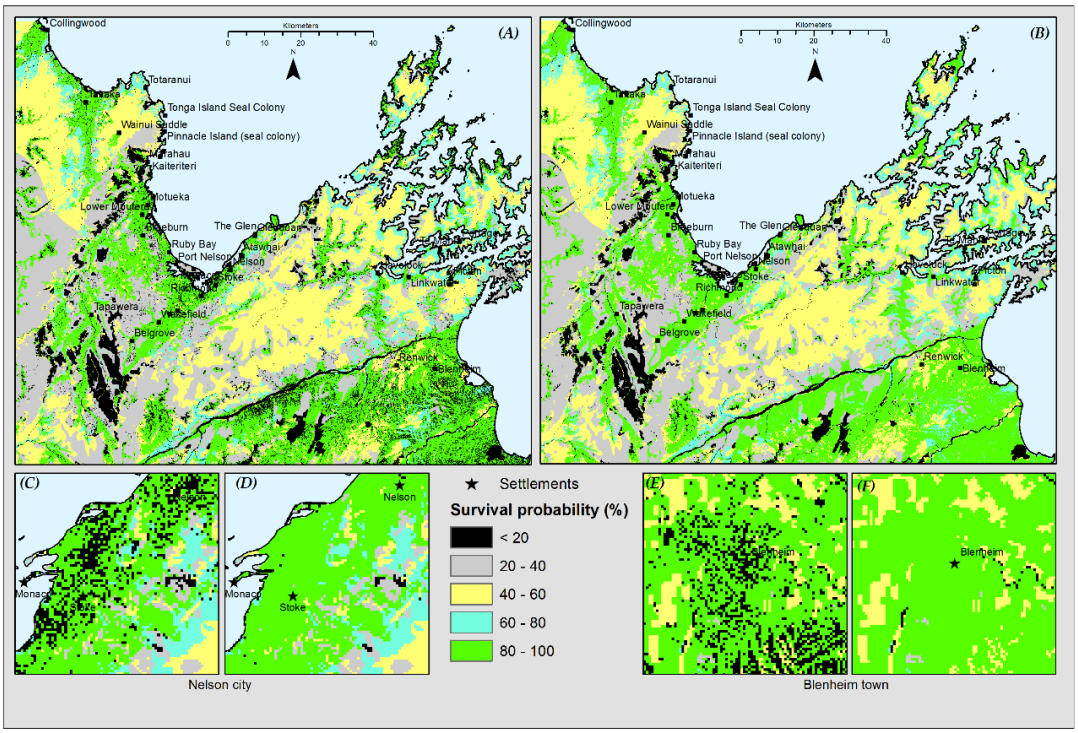


Figure 7.5: Survival layers used in the P. brassicae dispersal model.
(A) Survival layer 1(surv_1) composed of climate, land cover and growing degree day suitability layers with high resolution man-made structures layer used to increase geographic detail of urban areas. (B) Survival layer 2 (surv_2) composed of all suitability layers included in the first survival layer except the man-made structure layer from the high resolution satellite image. (C) & (D) show a close up map of Nelson city for Surv_1 and Surv_2 respectively, and (E) & (F) show a close up map of Blenheim town for Surv_1 and Surv_2 respectively.

7.2.6 Assessing the effect of eradication carried out during the years 2012-2013

The Department of Conservation's eradication programme against *P. brassicae* officially began in November 2012, but was not fully operational until early 2013. Mortality due to the eradication programme was simulated separately, then compared with the control simulation that used the survival layer, surv_1. The detection scheme given below was used in the field as a systematic strategy to eradicate *P. brassicae*, and the same method was used in the simulated eradication for the year 2012 and 2013 in this study. The detection and eradication team had delineated operational management blocks that encompass Nelson city and surrounding areas (Figure 7.6). The management blocks were generally classified as outlier areas, satellite areas and core areas. Four types of surveillance were carried out in

these management blocks. These were passive, active, general and follow-up surveillance (Phillips *et al.*, 2013).

Passive surveillance, was undertaken in response to reports of *P. brassicae* sightings by the public. Most of *P. brassicae* detection were recorded through passive surveillance.
Active surveillance is undertaken whenever there is a *P. brassicae* detection, including in response to a passive surveillance report of *P. brassicae*, and covers a pre-defined distance around the new detection. The radius of search depends on the management block the detection occurs in. Active surveillance is conducted within 200 m radius when the new detection is in the outlier management block, within 150 m when the detection is within the satellite areas block. If the new detection is within the core areas management block, no search is conducted as these areas are covered under the second surveillance scheme referred as the general surveillance. However, in case no general surveys is undertaken in that part of the core area, a 150 m radius is searched after new detection. In all cases of active surveillance if additional detection is encountered a new extended search radius is set up with the location of the new detection as the centre. This is repeatedly carried out until no new finds were made.
General surveillance refers to planned searches that are not in response to any positive finds, rather this strategy served as a reconnaissance survey to establish the general distribution of *P. brassicae* and make sure any *P. brassicae* eggs found do not progress to pupation. General surveillance was frequent in core areas, but was also carried out in historical hot spot areas in the satellite blocks that had been eradicated, to provide confidence that the invasion front did not expand into outlier areas.
Follow-up surveillance is undertaken when there is a specific visit to a property and *P. brassicae* was removed. This was designed to increase confidence other *P. brassicae* eggs, larvae or pupae do not remain in the area (Craig Phillips – Pers. Comm., 23 September 2013).

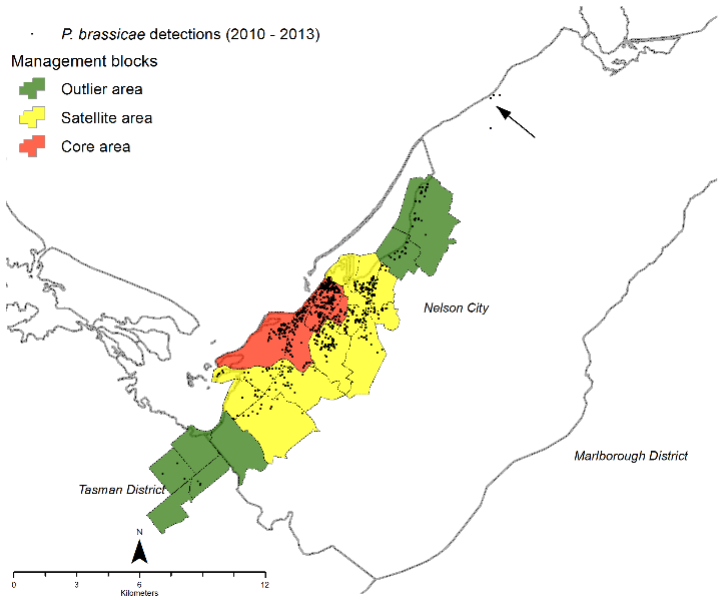


Figure 7.6: Eradication management blocks used by the department of conservation in Nelson city and surrounding areas.

The arrow indicates a number of P. brassicae *detections that are covered by surveillance in an outlier area management block. The block is not shown here as GIS data was not available. However, virtual eradications in that area were conducted as per the radii set for the outlier management block.*

Passive surveillance, was not simulated in this study; because it was undertaken in response to reports of *P. brassicae* sightings by the public, it could not have been replicated without the time and location of the public reports. Because, all the detections from follow-up and general surveillance methods were included in the occurrence data, they were treated with the same radii defined for search eradication under the active surveillance scheme (more description is given in Section 7.2.7).

7.2.7 Simulation

Sixteen years of simulations were undertaken representing dispersal from the year 2010, the first year of introduction of *P. brassicae* to New Zealand, to 2025. The simulation was replicated 1000 times to produce an occupancy envelope for dispersal at the end of each time step. Three thresholds [5, 10, 50] that corresponded to the number of times a cell was occupied during dispersal for all the replications was used to estimate probability of dispersal into a cell (Pitt *et al.*, 2009). A single site from which *P. brassicae* is thought to be introduced was used as an initial dispersal point in all the simulations.

According to the above simulation framework, three simulations were undertaken:

1) *Dispersal of* P. brassicae *on the Surv_1 landscape*
2) *Dispersal of* P. brassicae *on the Surv_2 landscape*
3) *Dispersal of* P. brassicae *on the Surv_1 landscape with a virtual eradication on year 2012 and 2013 using the detection and eradication buffers specified under the active surveillance method.*

The active surveillance details were used to apply the virtual eradication on the simulated dispersal maps of year 2012 and 2013. Eradication buffers were constructed according to the specifications of the active surveillance radii for the different management blocks. The different buffers representing the different radii within different management blocks were then merged to create an area of virtual eradication (Figure 7.7). All dispersers within the buffer of eradicated areas in 2012 were removed from the year 2012 simulation before simulating the dispersal for the next year. This was also carried out for the year 2013. From

2013 onwards the dispersal was simulated until year 2025 without any eradication. The dispersal pattern, area and final state was compared with the control simulation that had no eradication (simulation with Surv_1).

The simulation was run on a quad core PC with 3.40GHz Intel i7 processors, and each simulation of 16 years by 1000 replicates took 32.5 hours. Total area covered at each time step by the predicted *P. brassicae* dispersal was compared for dispersal generated based on surv_1 and surv_2 survival layers, and for simulation based on Surv_1 with eradication.

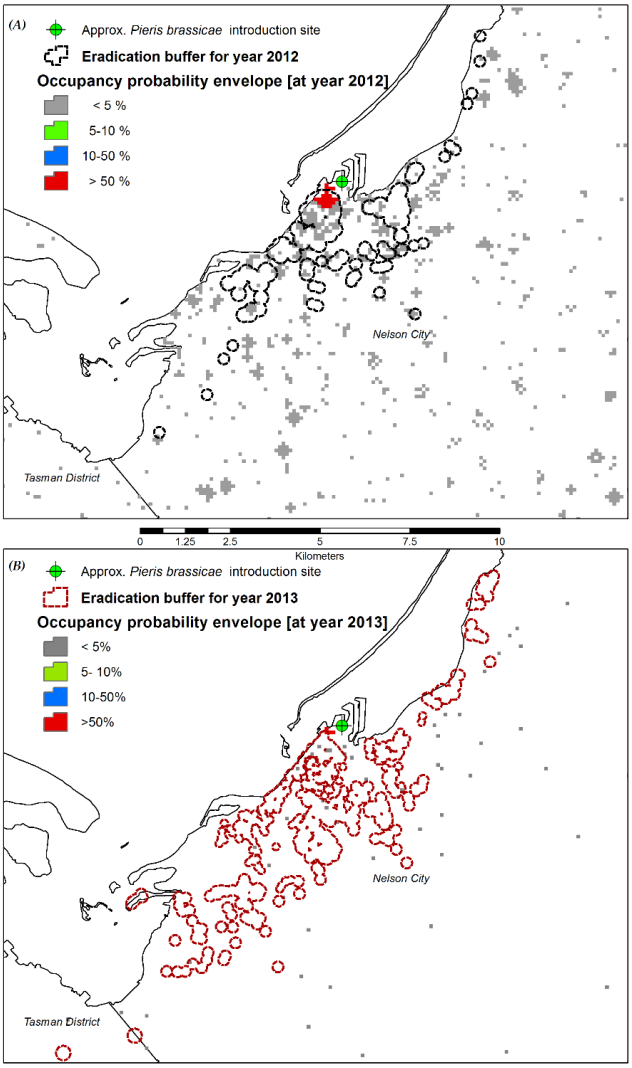


Figure 7.7: Virtual eradication buffers specified according to the active surveillance –eradication scheme currently used on the ground in the years (A) 2012 and (B) 2013
The occupancy probability envelopes are shown to give the overall view of the dispersal at the year 2012 & 2013 over the 1000 simulations. The virtual eradiation was carried out om each of the replicates separately. Because, a separate detection method was not designed, the buffers were drawn around the real detections for each year and were applied on the simulated dispersal.

The New Zealand data set of *P. brassicae* detections and absences obtained from the Department of Conservation (Phillips *et al.*, 2013). were used to compare the first three years of the dispersal occupancy envelopes of both Surv_1 and Surv_2 dispersal model outputs with field data. Even though the original data was collated with higher temporal resolution, the data is presented here by year as the simulation is set up for yearly parameters.

Table 7.3: Geographically refernced *P. brassicae* survey data points available for validation

Year	Presence	Absence	Total
2010	19	-	19
2011	26	-	26
2012	199	614	813
2013	835	26498	27333

7.3 Results

7.3.1 Dispersal parameters

The estimated dispersal parameters based on *P. brassicae* occurrence points in the United Kingdom were used to calibrate the kernel module to generate long distance dispersals (Table 7.3). Parameters estimated using New Zealand spatio-temporal occurrence records are also reported just for comparison, they were not used to calibrate the model. However, the limited occurrence points in New Zealand were used to validate the simulated eradication.

Dispersal parameters were extracted based on the distances between successive dispersal events and their frequency. The *P. brassicae* dispersal events in the United Kingdom showed a typical fat tailed distribution where most dispersals were near existing sites, with a few rare dispersals located further from the invading front (Figure 7.8-A). Such rare long distance events however were found to be essential for the rapid advancement of an invading species as these colonisers could disperse even more locally where numerous unoccupied sites could be accessed compared to sites that are within the main invasion front (Higgins & Richardson, 1999; Cain *et al.*, 2000).

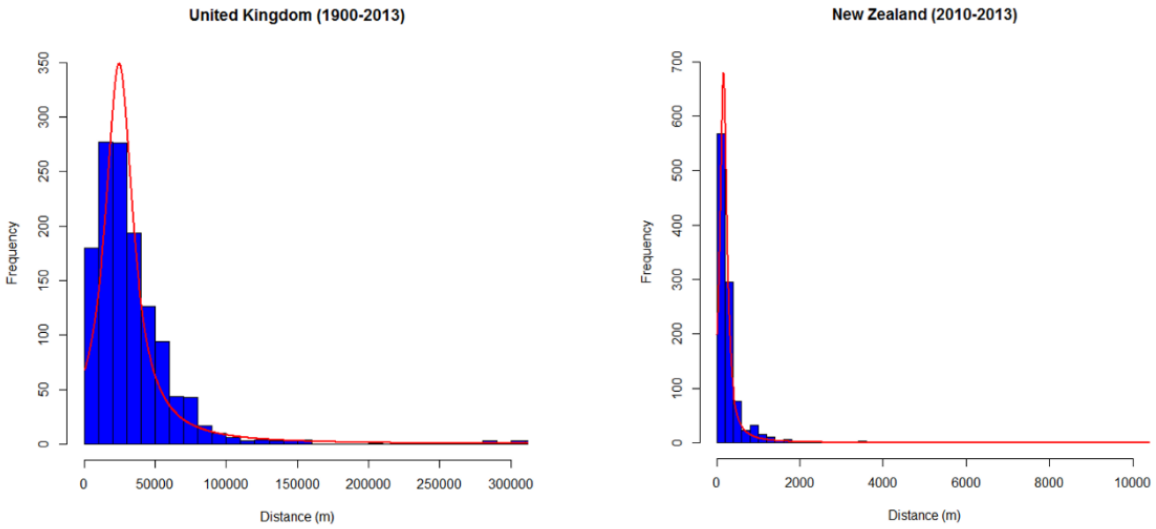


Figure 7.8: distances between pre-existing and newly occupied P. brassicae *sites fitted to the Cauchy probability density function of the user defined parameter values for United Kingdom data (A), and New Zealand data (B).*

The New Zealand dispersal distances data was not used to calibrate the model however the graph is shown here (Figure 7.8-B) to compare the result with the dispersal distances in the United Kingdom. Even though *P. brassicae* was only introduced in New Zealand in 2010 a similar trend of a stratified dispersal can be seen from the dispersal distances.

The error introduced due to the digitization of the occurrence data from the scanned and rectified image was estimated at ± 146 m. The distance parameter estimation of the shape and scale parameters of the Cauchy distribution was averaged from a thousand runs of noised data to account for the uncertainty from the digitized occurrence points in the UK. The estimated values from the un-noised data (Figure 7.9) had very little variation to the average of the noised data and the standard deviation as a whole was very low (Table 7.4.). This was expected as the survey points in the UK were representative of 10 km^2 which was a larger area than the estimated uncertainty introduced from the digitized points (145 m^2). Additionally, as explained by Pitt *et al.* (2011), the parameters were more affected by variation in the tail of the Cauchy distribution, where the distances associated with this part of the distribution are very large, which requires high location uncertainty to affect the parameters.

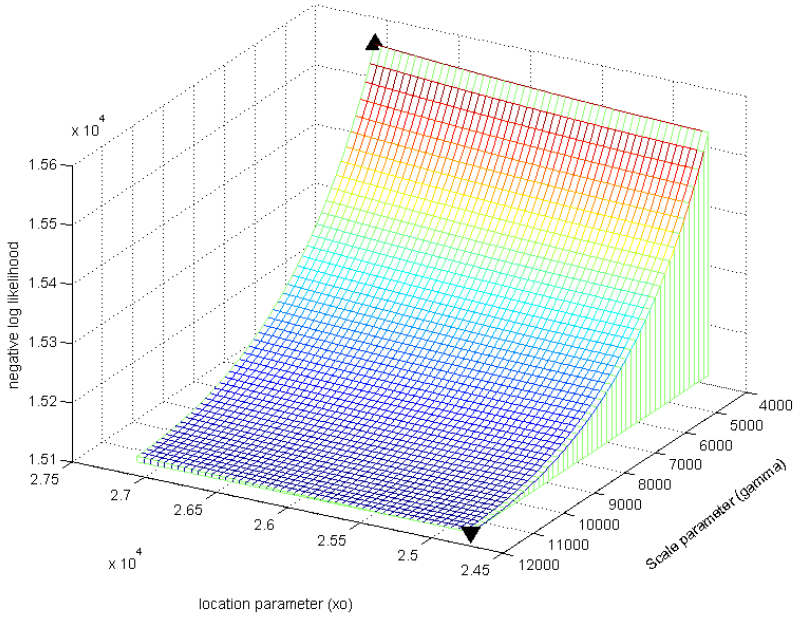


Figure 7.9: Maximum likelihood iterations to optimize distance parameter estimation. Black triangles indicate the initial values at the top and the best values at the bottom.

The frequency of dispersal events was more in agreement between the two datasets than the distance of the events. The average number of events that were estimated to occur per each existing site was 0.41 in case of the United Kingdom dataset and 0.48 in case of the New Zealand dataset. The estimated distance (x_o, γ) and frequency (λ) parameters for both United Kingdom and New Zealand are given in Table 6.3.

Table 7.4: Dispersal parameter estimates used to calibrate the MDiG model for the *P. brassicae* dispersal simulation.

Data	Cauchy - x_o (location)	Cauchy - γ (scale)	Poisson – λ (expected mean)
United Kingdom	24 752 ± 5.22	11 894 ± 3.23	0.41 [LCI=0.02, UCI=1.99, α=0.05]
New Zealand	154.22	99.43	0.48 [LCI=0.05, UCI=1.77, α=0.05]

** The Xo and γ estimates were used to characterise the Cauchy distribution used by the kernel module to determine dispersal distances for the stochastic dispersal events generated from all occupied cells. The λ parameter was used to characterise the Poisson distribution for the kernel module to determine the number of dispersal events generated from occupied cells. The direction of the generated dispersal events was not user parameterised and the MDiG (Pitt et al., 2009) default setting of random selection of directions between the range of 0 - 2π radians from a uniform distribution was used.*

7.3.2 Comparison of occupancy envelopes

7.3.2.1 Comparison based on field data (year 2010 – 2013)

The occupancy envelopes from both dispersal simulations based on the two survival layers described in the method section were validated using the New Zealand *P. brassicae* temporal occurrence data obtained from DOC. Even though the New Zealand invasion of *P. brassicae* was well surveyed from the time it was detected in 2010, insufficient years have elapsed to use this data for validating different occupancy thresholds. This is because initial dispersal simulations differ due to stochasticity of the dispersal process (Pitt *et al.*, 2009).

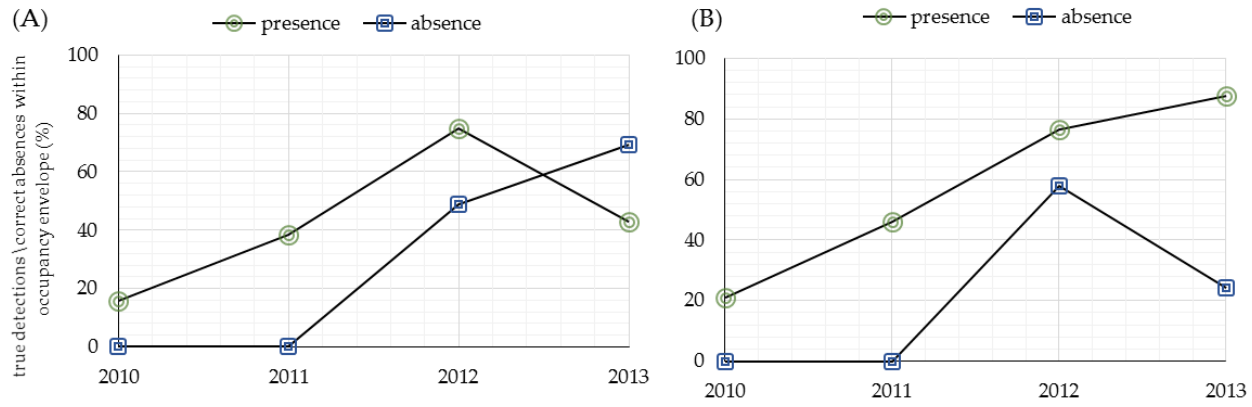


Figure 7.10: Percentage of correctly represented presences/absences by the occupancy envelopes of dispersal model with the first survival layer (A) dispersal model with the second survival layer (B).

In my study higher threshold percentage occupancies were not reached until the simulation year 2015, therefore the "> 0 %" threshold was used to evaluate the dispersal outputs. The unique opportunity gained from this validation dataset was that absence data was collected during *P. brassicae* surveillance, which allowed me to assess the mean yearly specificity of the dispersal model. Accurate temporal absence data are often very difficult to obtain because detections often rely on passive finds, where only positive sites are recorded as opposed to sites where the species was not present.

The percentage of presence sites that were correctly contained within the occupancy envelopes increased for both dispersal models with Surv_1 and Surv_2 layers from year 2010-2012 (Figure 7.10). However, the percentage decreased for the model with a survival layer of added geographic detail at urban areas (Surv_1) for the year 2012-2013, while the

model with the Surv_2 layer increased. For absence sites that were correctly left out of the occupancy envelopes, the model using Surv_1 layer showed an increasing trend from 2012 to 2013 (two years when absence data were available), but the model based on the survival layer Surv_2 showed a steep decline in the number of correct absences.

Three measures were used to estimate the mean yearly performance of the dispersal models (Table 7.5). Even though the Surv_2 model scored more correctly predicted presence sites for the year 2010-2013, further examination of its accuracy and specificity showed it was not a superior model.

Table 7.5: Performance scores for the dispersal model based on two different survival layers

Scores (%)	*Surv_1*	*Surv_2*	*Remark*
Accuracy	68.07	27.27	$\frac{TP+TN}{TP+TN+FP+FN}$
Sensitivity	48.29	83.31	$\frac{TP}{TP+FN}$
Specificity	68.86	25.04	$\frac{TN}{TN+FP}$

TP=True positive, TN=True negative, FP=False positive and FN=False negative

The Surv_1 model gave higher overall mean yearly accuracy and specificity (about three times that of Surv_2), but a lower sensitivity score. This could be due to inaccuracy from dispersal parameters. But more importantly it could be the result of up scaling of the 2.5m resolution man-made structure layer from the remotely sensed image to a 100m resolution survival dataset. The up-scaling may have resulted in some garden and park spaces being included within the man-made structure layer. If the latter is the case, it will only affect urban locations, as the man-made structure layer was used to re-code areas of the landscape where there are man-made structures. The high sensitivity obtained from model Surv_2 was due to the high survival value given to all urban areas. However, unsuitable sites were also incorrectly labelled as suitable, so specificity was low.

Both models closely simulated the progression of the Nelson inner city invasion as per the pattern observed from the occurrence data except for the first simulation year (2011). For 2011, there were more actual occurrences compared to the simulated dispersal. That is clearly due to not enough source cells when the dispersal phase started. We assume that *P.*

brassicae was first introduced into New Zealand in 2010 based on the year of detection. (Reference map for place names in the study area are given in Appendix 7.4)

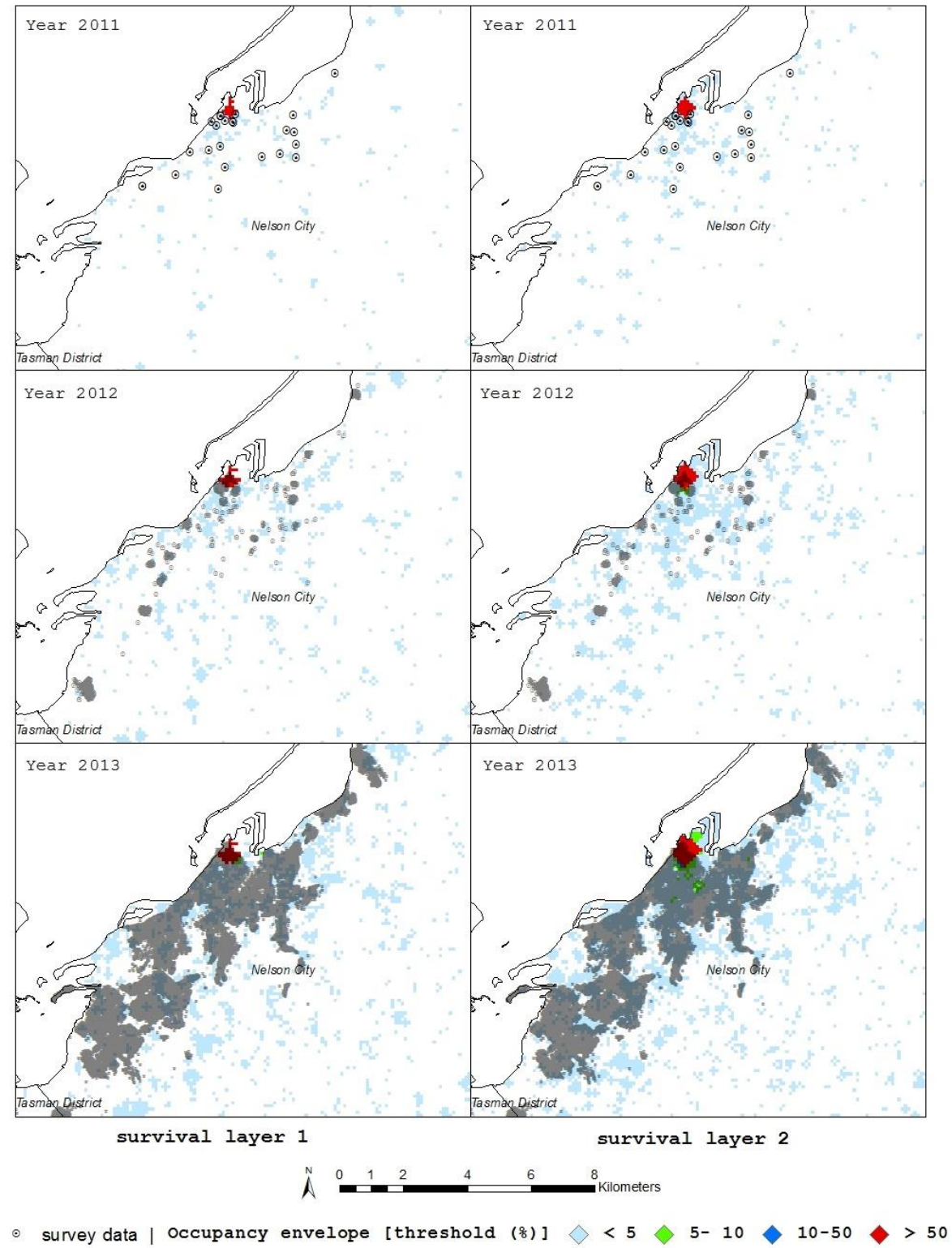


Figure 7.11: Dispersal maps overlaid with P. brassicae *presences from field survey data*

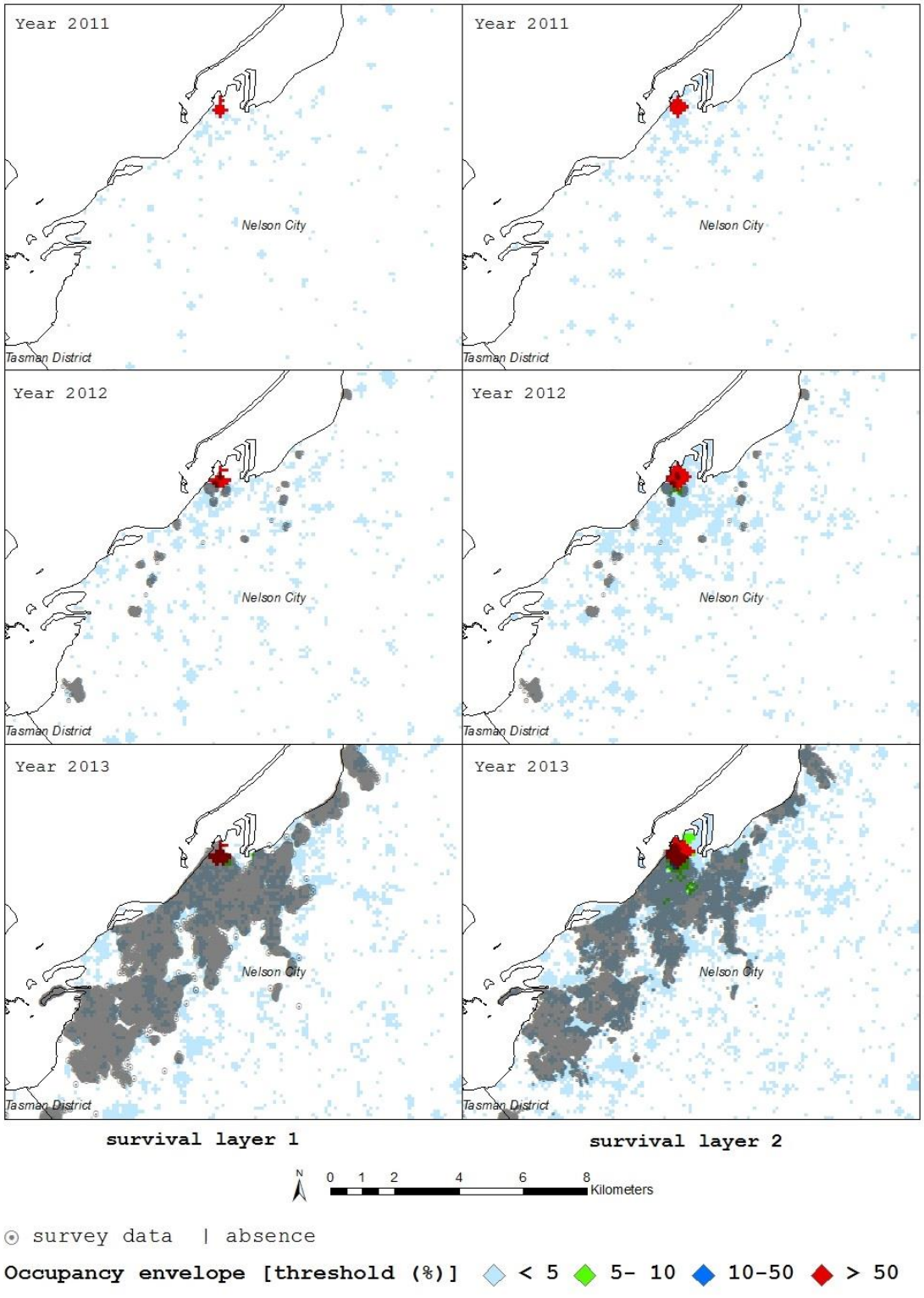


Figure 7.12: Dispersal maps overlaid with P. brassicae *absences from field survey data*
Note: There was no absence data available for the year 2011.

However, it is quite possible that the species might have completed a generation or two before it was detected, which might explain the more dispersed surveillance data compared

to the few cells predicted by the models for the year 2011. Most importantly, the discrepancy could result because it is difficult to approximate realistic dispersal patterns early on the simulation.

In general The Surv_2 model contained more absence areas within the occupancy envelope than the Surv_1 model. When comparing the number of satellite populations that were correctly predicted through 2010-2013, both models predicted the satellite populations detected in Atawhai with the Surv_1 model being more spatially precise. In 2012 both models predicted the village Stoke where another population was detected further from the front close to port of Nelson. However, with the Surv_2 model more spatially precise (Figure 7.11). In 2013, both models predicted the satellite population found in the towns of Richmond and Atawhai.

7.3.2.2 Area covered (year 2010 – 2025)

Different probability thresholds [5, 10, 50] were used to analyse the area covered at different time steps of the simulation. Accordingly, four probability of < 5%, 5-10%, 10 -50 % and > 50 % occupancy envelopes were assessed. The area of the "<5%" occupancy envelope has increased until around 2018 and 2019 in case of Surv_2 and Surv_1 models respectively before declining consistently until the end of the simulation year 2025. This trend shows less agreement among replicates in the beginning of simulation which is the effect of the stochastic nature of the dispersal. However as the years pass the effect of the landscape shaping dispersal becomes apparent with more replicates agreeing, hence the increase of the area of the ">50 %" envelope right after years 2022 and 2021 for Surv_1 and Surv_2 dispersal models respectively. The time gap with respect to the filling in the higher percentage occupancy envelopes between the two models shows delayed dispersal caused by higher percentage unsuitable cells within highly suitable neighbourhoods for the Surv_1 model.

The exponential relationship between dispersal events and the number of occupied sites was truncated due to small study area extent forming a logarithmic relationship as shown in the case of both dispersal models. The model with Surv_2 layer however (Figure 7.13-B) reached equilibrium more quickly, i.e. all suitable areas were saturated.

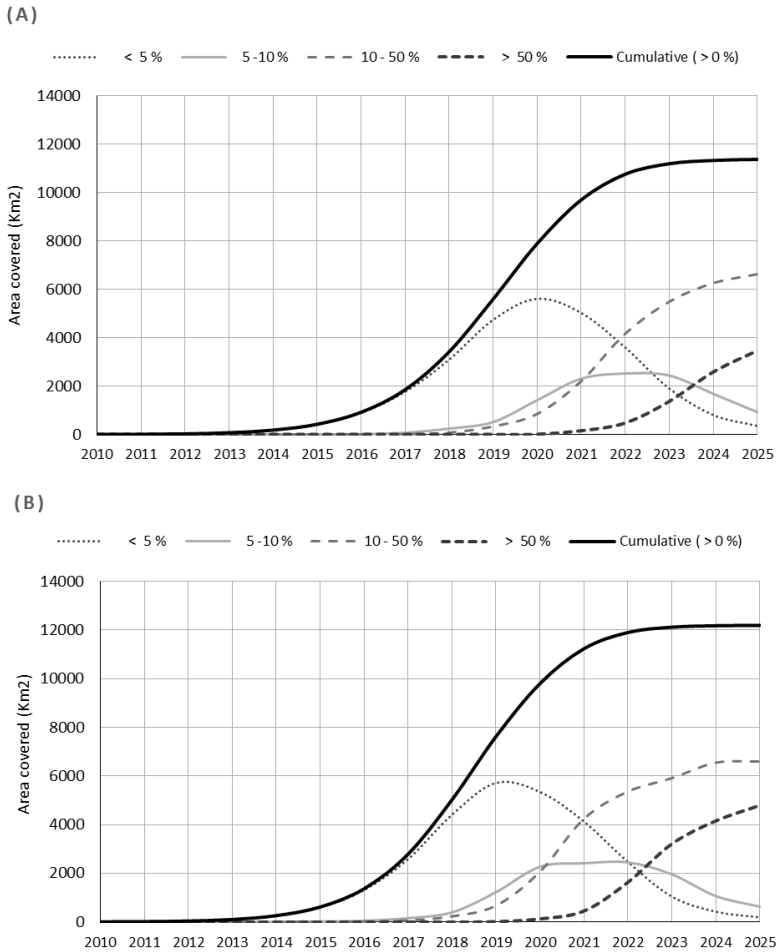


Figure 7.13: Area covered by the simulation occurrence envelopes for various thresholds. (A) Dispersal using the first survival layer with added geographic detail from satellite images in urban areas, (B) survival layer without the higher resolution geographic detail in urban areas.

The comparison between the cumulative probability envelopes for threshold > 0% (accounting for all sites predicted at least once) of the two dispersal models showed that more dispersal sites were covered by the model using Surv_2 at all of the time steps (Figure 7.14-A). This is expected as the first survival layer (Surv_1) with more unsuitable cells had fewer suitable cells compared to the surv_2 layer because man-made structures within urban areas were set to a survival probability of zero. However, when the difference in predicted dispersal area was weighted with the number of high survival cells (survival probability > 90%), for both models, there was still a difference in the area covered by the dispersal as well as fluctuations in the increase in area over time (Figure 7.14-B). This suggests that the higher precision in mapping unsuitable patches among highly suitable areas slowed dispersal, hence, lowering the rate at which suitable cells were occupied in

addition to the cells that could not be occupied simply because they were recoded to zero survival probability.

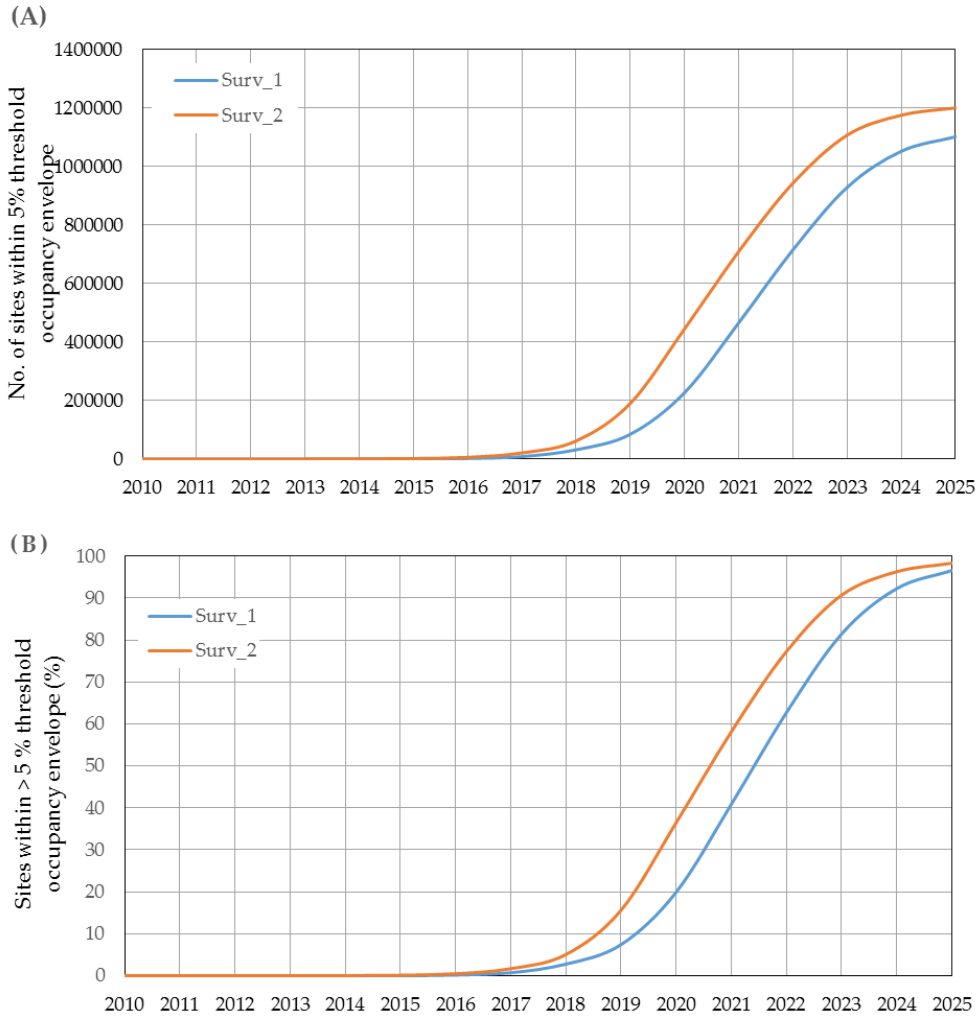


Figure 7.14: Comparison between the dispersal models using the survival layers tested in this study. (A) The number of sites covered under the "> 5%" occupancy envelope threshold. (B) The percentage of area covered by the ">5%" envelope by the two models after correcting for the difference in the available high survival probability sites.

The greater spatial detail within highly suitable areas as in Surv_1, meant local dispersal was slowed as dispersers took more time to spread through a neighbourhood of highly variable survival probability than the time it takes to spread through a neighbourhood of uniformly high survival probability.

When the actual dispersal maps are compared there is an apparent delay in occupation of suitable areas when the first survival layer (Surv_1) is used. For example, by 2020 the Surv_2

dispersal envelope of higher thresholds reached Renwick and Blenheim and covered extensive areas beyond the Wairau valley in the Marlborough district. Whereas, the Surv_1 dispersal model did not have any high percentage threshold envelopes that reached Renwick or Blenheim and only limited dispersal within the 10 -50 % envelope reached beyond Wairau valley (Figure 7.15). It is also notable that by the end of 2025 the Surv_2 dispersal model > 50% threshold envelope, covered extensive areas in the bays, islands and peninsulas of Marlborough Sounds, while these areas were still not covered by the high percentage envelope generated by Surv_1 (Figure 7.15).

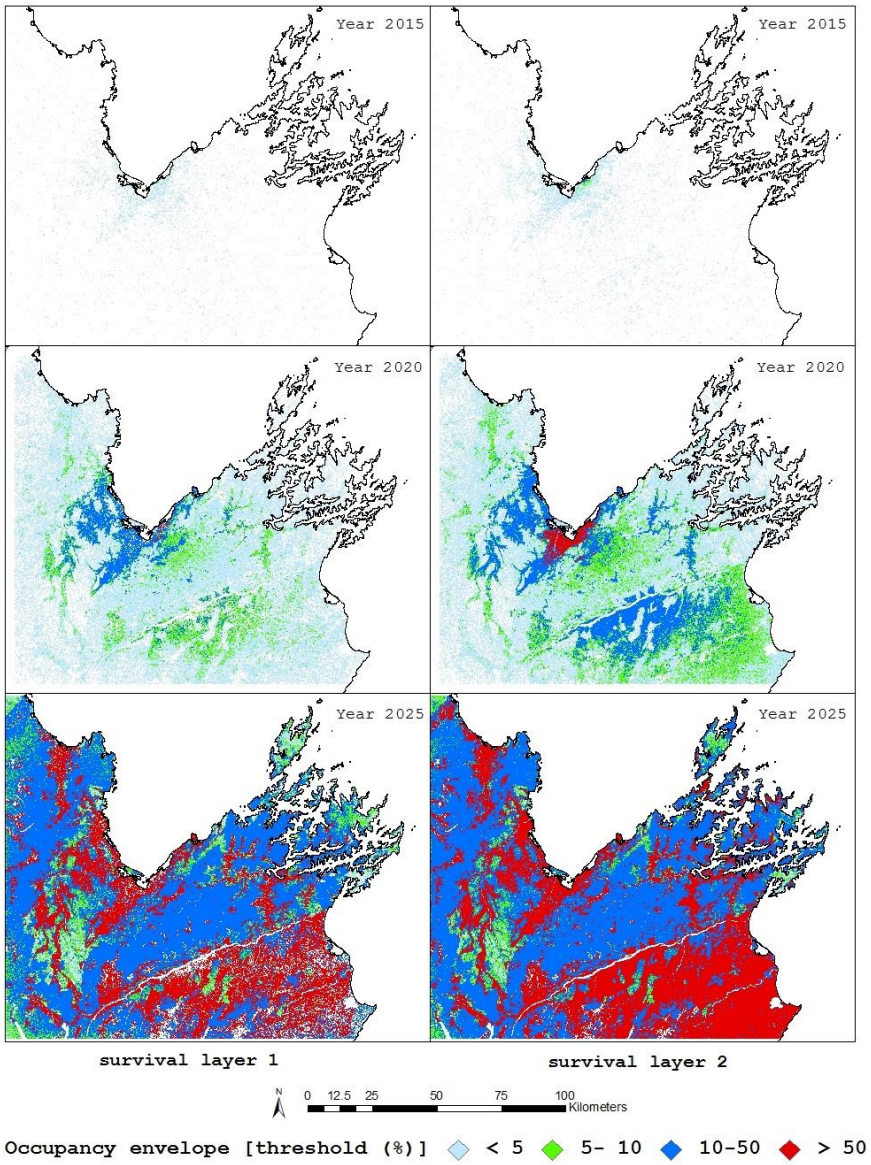


Figure 7.15: Dispersal coverage for the year 2015, 2020 and 2025 based on survival layer one (left panel) and survival layer two (Right panel)

7.3.3 Effects of current eradication scheme on dispersal dynamics

The dispersal area covered by Surv_1 (the control dispersal model with no eradication) and the model with a portion of dispersers virtually eradicated on year 2012 and 2013 were compared.

The difference in area covered showed that the eradication successfully suppressed long distance dispersal for four years (until 2016), before it slowly starts to climb steadily to 2025 (Figure 7.16). At the end of the simulation it is apparent that the eradication model does not reach equilibrium with large areas still unoccupied in 2025 (Figure 7.17).

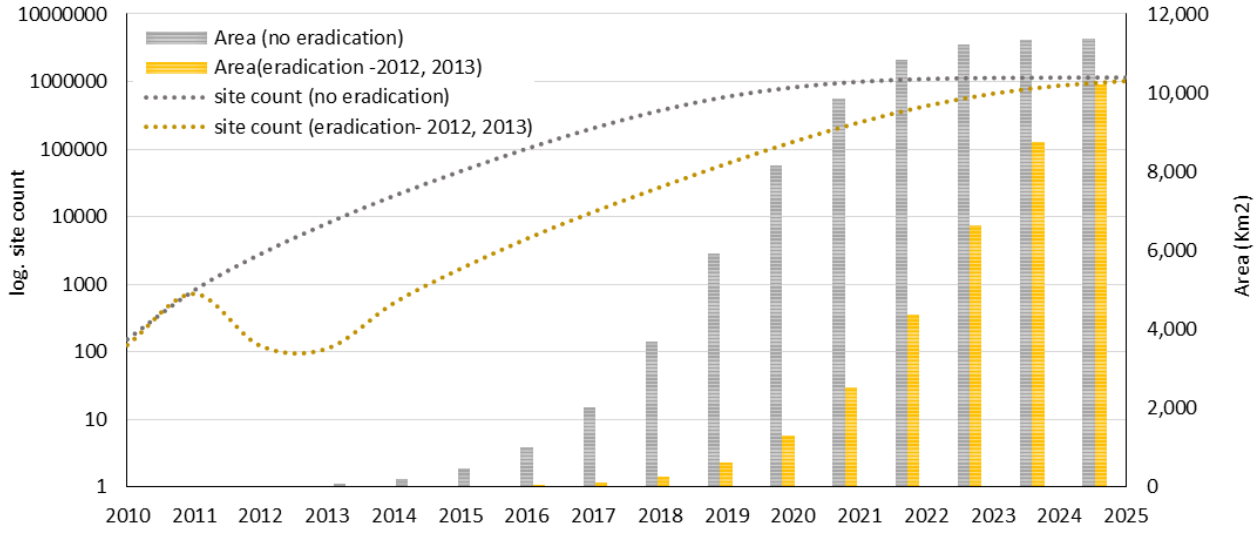


Figure 7.16: Comparison between the control dispersal model and the eradication model replicating the current eradication scheme. Left axis: logarithm of site counts, Right axis: Area (km²)
The site count and area comparisons given for the control model and the eradication model based on the > 0 % probability threshold which includes all sites that were predicted at least once within the occupancy envelop of 1000 replicates.

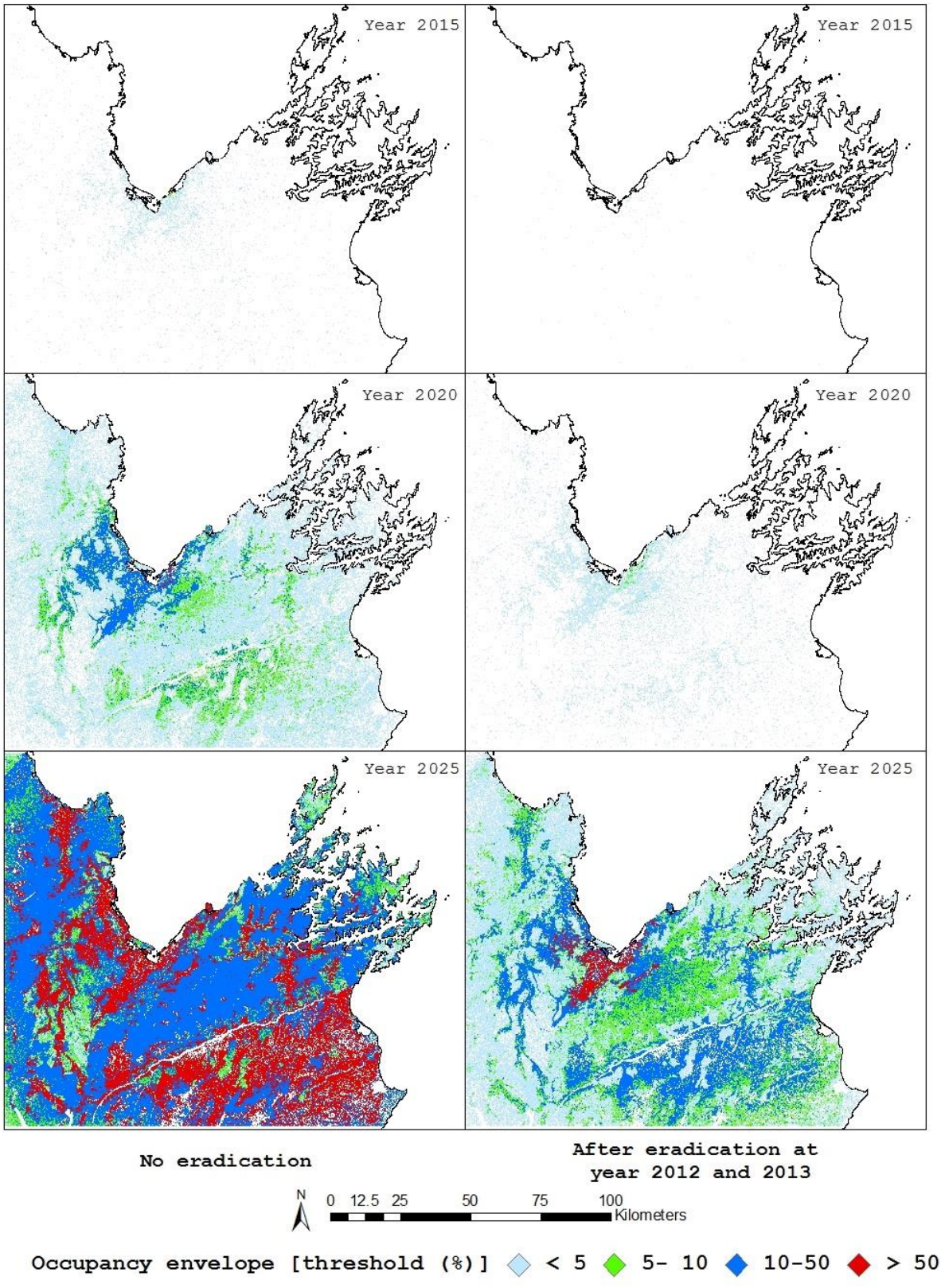


Figure 7.17: Dispersal coverage for the year 2015, 2020 and 2025 based on survival layer 1 with no virtual eradication (left panel) and after eradication in the years 2012 and 2013 (Right panel)

7.4 Discussion

Mapping and predicting species distributions is subject to uncertainty due to insufficient biotic and abiotic data (Pearson *et al.*, 2006); predicting dispersal in a spatio-temporally explicit manner is even more uncertain (Dunning Jr *et al.*, 1995). This is because uncertainties from early time steps get compounded into subsequent time steps, which adds error to estimates of both space and time (Pitt *et al.*, 2011). However the increasing threat from invasive species and the need for integrated surveillance and eradication schemes calls for some sort of decisions and ideally, input from species distribution and dispersal models. The alternative is to cover all possible areas the species under surveillance could invade. That often leads to a program that is far too costly and unmanageable (Hulme, 2006).

In this study, the simulation time frame was limited to 16 years to keep any dispersal pattern or rate predictions within a reasonable time frame that are only likely to be affected by the initial conditions and dispersal parameters provided. While discussing the temporal limitations of species dispersal models, Pitt *et al.* (2011) argued that over time, dispersal realizations at future time steps begin deviating from the pattern that is associated with initial conditions because of increasing uncertainties that build up over the years of the simulations to a point where the realizations will no longer be informative. In part, such uncertainty results from the inherent random nature of the long-distance kernel used in the dispersal model. While there is an alternative to use a deterministic model that estimates distance and frequency of dispersal based on pre-set conditions or landscape types, such kernels usually require extensive parameterization that is only suitable for a species that has been studied over a long time rather than a newly arriving species about which information is limited.

Regarding spatial extent, the modelling framework was kept within a limited spatial extent (12,466 km^2) so that it is easy to understand or interpret the dispersal pattern once the species reaches equilibrium with respect to the available area. The limited extent gives insight into which unsuitable areas get occupied due to propagule pressure despite having a relatively lesser survival probability assigned in the model. Such understanding of the dispersal dynamics is important from the management aspect where vulnerable areas even

within a seemingly unsuitable environment can be identified. A good example is the recurrent occupation of Rabbit Island despite having quite unsuitable land cover. In such cases, one might expect a transient population that can only populate the area in single time-steps without having a permanent population. Such, interpretation of dispersal results can be used to prioritize a within border quarantine system so that invasive species can be deterred from spreading further within a country as discussed by Paini *et al.* (2010) who suggest that management of invasive species within borders should be part of a larger biosecurity framework.

7.4.1 Improved representation of species-landscape interaction

It is already been well established that the heterogeneous landscape affects species distribution (Dunning Jr *et al.*, 1995; Higgins *et al.*, 1996; Pitt *et al.*, 2009) and that species dispersal realizations at later time steps of dispersal simulations usually follow the landscape pattern (Pitt *et al.*, 2009). Despite the fact that it is difficult to get high confidence in terms of spatial specificity from stochastic dispersal models such as used in this study, Gilbert *et al.* (2005) suggested that it should be possible to achieve a certain level of confidence in the spatially explicit results of such models, if precise landscape data that specifies areas where the landscape could exacerbate and/or limit species mobility are used.

Since accurate landscape information is important for spatially explicit dispersal models (Baker, 1993), the next step is to work out an optimum resolution as well as an optimum level of detail to represent the area of interest for dispersal studies. Recommending generalized optimum landscape measures is outside the scope of this study as it requires extensive research design involving factorial experiments to test multiple landscape types (Saura & Martinez-Millan, 2001). Instead, what is shown in this study is a method where the major part of the landscape is constructed from a lower resolution data with specific areas enhanced with high resolution data.

Recoding the homogeneous urban area was necessary (as done in Surv_1), as otherwise the whole area would have to be labelled as suitable (Surv_2 model) to facilitate dispersal through urban areas. Because there would be no unsuitable patches in such an area the dispersal predictions would overestimate the dispersal rate throughout urban areas

providing more occupied cells that could give rise to multiple long distance dispersal. The most important implication of such a result is that overestimating future dispersal might incorrectly discourage an eradication attempt.

There are many situations where selective recoding of the landscape could be applied. For example, in the case of *P. brassicae*, only the geographic detail in urban areas that highlights unsuitable manmade structures was needed therefore a visible band image was used. However, if the target species was a forest pest with a specific host, and a diverse forest landscape is available, then, keeping all other land cover data constant, a medium resolution but hyper-spectral image of forest areas could be used to map the particular host species by giving host trees higher survival probability than other tree species. The result will be more accurate dispersal estimation through forest areas compared with labelling all forest cover with high survival probability.

7.4.2 Dispersal parameters and model choice

Prediction science often considers that the best model is the most parsimonious. In such a study as this one, that would mean the model which can explain most of the species spread and dispersal dynamics while allowing a high level of abstraction (Andow *et al.*, 1990). According to this study, it is extremely difficult to parameterize species dispersal even when working with a species that has been well studied in the past, such as *P. brassicae*. The difficulty of parameterising species dispersal is a view also shared by many important dispersal researchers (Shigesada *et al.*, 1995; Higgins & Richardson, 1999; Cain *et al.*, 2000). The difficulty is due to the countless alternatives a disperser can take especially if it uses a combination of both active and passive dispersal and the inherent stochastic nature of species dispersal in general. Parameter estimation becomes even more complicated when known dispersal mechanisms can manifest a different pattern when realized over a new landscape. Nonetheless, no matter what model is used (either data driven or expert driven) we try to characterize the most likely mode of dispersal and behaviour of the target species which allows the pattern and rate of a new invasion to be approximately predicted.

The particular advantage of using MDiG is the relatively simple assumption associated with the kernel module that represents long distance dispersal. The kernel represents the

outcome of a complicated interplay of factors that are not represented individually in the module but are characterised by an overall pattern formed as a result of these factors. This assumption simplified the long distance dispersal estimation process while explaining the dispersal pattern of the target species if the distance and frequency parameters are obtained from a carefully analysed empirical dispersal data.

Pitt *et al.* (2009) found the MDiG simulation model did better in later time steps than a radial diffusion model but was weaker in earlier time steps. Savage and Renton (2014) called for a generalized model for biological invasion, and suggested, that MDiG is the closest to a generalized model for biological invasions but remarked that " [M]DiG is more of a regional model that simulates spread over many years over regional or national scale". In my opinion both of the above statements may have overlooked the role of the powerful "neighbourhood" module to fine tune spread at local scale, which means MDiG may have increased precision in capturing initial dispersal conditions and early spread as well as its effective portrayal of invasions at a later stage of simulation. However, further studies need to be conducted to confirm this assertion.

Finally, it is important to highlight that in this study the dispersal parameters x_o, γ and λ were assumed to represent a combination of different factors that derive *P. brassicae* long distance dispersal. As well, while great care was taken to use a similar source landscape and representative historical data to estimate the parameters, it is possible different factors might be important in the study area in New Zealand compared to the source area in the United Kingdom from which the parameters were estimated. Testing variation around the parameters instead of using the mean as done in this study would increase understanding about the effect of the parameter error on dispersal and help fine tune simulations (Pitt *et al.*, 2009). More important is the need to re-calibrate dispersal models as more field information becomes available to minimize uncertainty in dispersal outputs that may be compounded as a result of propagation of error through time.

7.4.3 The future of **P. brassicae** *in New Zealand*

The dispersal pattern of *P. brassicae* according to the models used in this study is characterised by a rapidly moving front that has numerous advancing satellite populations.

When these satellite populations are not too far from the advancing front the rate of total area covered by the main front increased significantly as individuals in the satellite population disperse back into the main front. The *P. brassicae* dispersal pattern can be described as a short rotation of advance and coalesce (Shigesada *et al.*, 1995) sequence as observed from the occupancy envelopes as well as the actual surveillance data collected by DOC from 2010 to 2013.

Previous studies have reported that a female *P. brassicae* adult embarks on a long distance flight to look for a suitable host if the current site is overloaded with larvae. In the case of *P. brassicae* long distance dispersal can therefore be density dependent where the density of larvae on host plants provide the stimulus to initiate active flights. The dispersal scheme used in this model represents this process. Long distance events were generated from existing or occupied sites in the earlier time steps compared with long distance events occurring independently from already occupied sites as in Robinet *et al.* (2009). This relationship between occupied sites and future dispersal events can readily be demonstrated by the suppression of dispersal events due to the intervention undertaken in 2012 and 2013, where the removal of possible dispersal source sites set back the dispersal a few years when compared to the control dispersal model. A report by Phillips *et al.* (2013) showed that the current eradication scheme was successful in suppressing *P. brassicae* outliers in Glenduan and Richmond; if this trend extends for all populations of *P. brassicae* in and around Nelson, it may facilitate the elimination of *P. brassicae* from New Zealand (Phillips *et al.*, 2013).

The timing of the intervention by DOC was probably the most critical factor that led to the predicted suppression of *P. brassicae* dispersal in the simulation. Based on the phenological model by Kean and Phillips (2013c), 2-3 *P. brassicae* generations per year are expected in Nelson. The intervention was started within two years of first detection of *P. brassicae.* That means only about *P. brassicae* 4-6 generations elapsed after its introduction, which meant greater success for possible eradication or containment. Any adult *P. brassicae* that were not eradicated in 2013 would likely be laying eggs back in the core invaded area, as it will not be overburdened by larvae and there would be enough host plants as a result of the earlier eradication, eliminating the need for density dependent long distance active flights. It is also

important to note only the eradication that was preceded by active surveillance was assessed in this study and the effect from the intervention could have been be more pronounced if an eradication as a result of the follow-up, and passive surveillance, were considered.

Eradication in the core area and elimination of satellites is shown to suppress the dispersal of *P. brassicae* according to the virtual eradication simulation (Section 7.3.3) based on the assessment of the current surveillance and eradication scheme. However, because external factors like parasitism and competition were not included in this study, further study that investigates alternative optimum detection and eradication schemes is necessary to ascertain the most efficient intervention.

There are a number of factors that are unaccounted for in this study. One is further introduction of additional *P. brassicae* to New Zealand. This is a very important factor as any success in *P. brassicae* eradication depends on keeping the density of the invading population to a minimum. Any top up from undiscovered introduction pathways could compromise the eradication effort. This is especially important for *P. brassicae* as it is known to effectively synchronize mating between generations from different source populations and that is the reason for its persistence in some parts of the British Isles as well as most of its range above the Arctic Circle (Feltwell, 1982). Even if this particular introduction into New Zealand might be a one off accidental incursion, identifying the pathway is important to prevent recurring introductions.

The other important factor is identification of possible hosts that might sustain *P. brassicae* in the event of heavy eradication campaigns in home gardens and farm lands. The native New Zealand cress species has already been identified as a possible target but more studies need to be undertaken to quantify possible alternate hosts of *P. brassicae* in the region. For example, the invasive sea rocket (both *Cakile maritima* and *Cakile edentula*) which was first reported in New Zealand in the 1940's (Cousens & Cousens, 2011) is also known to be an alternate host for *P. brassicae* (Feltwell, 1982) and could play an important role in the spread of *P. brassicae*.

7.5 Summary

Developing realised dispersal models is critical to inform a biosecurity eradication and surveillance team. Pitt *et al.* (2009) and Worner (1994) mentioned that dispersal models need to be realistically designed to represent the heterogeneous landscape. Realistic landscape representation of invaded ranges increases the accuracy of dispersal model results delivering precise information for policy makers and technical taskforces assigned to deal with invasive species surveillance and eradication (Dunning Jr *et al.*, 1995).

However, the optimum spatial extent, resolution and composition (level of detail) one should use for species dispersal models has usually been dependent on the species modelled, their dispersal pathways and the available data (Andow *et al.*, 1990). Essentially it is not possible to recommend a set of optimum landscape characteristics that could be generalized across multiple species due to multiple variations both in landscapes and in the dispersal behaviour among species (Levin, 1992). The approach of using a selectively re-coded landscape in this study gives the option of focusing on areas where the landscape has the most effect on dispersal dynamics enabling users to spend less time and money on areas that do not require high resolution data.

Even though *P. brassicae* is in the early stage of its invasion in New Zealand and the pattern observed from the surveillance data could be less dependable for this reason, the closeness of the simulation results to the observed invasion pattern of *P. brassicae* in Nelson is a good indication that such simulation based dispersal models could be very useful. Moreover, careful and detailed accounting of assumptions as well as estimation of dispersal parameters are important to replicate the dispersal pattern and dynamics of an invasive species which can be used to draw useful conclusions regarding future regional movements of the species. Last but not least, it is very important to re-calibrate dispersal models with current dispersal data which can be used to check for deviations from the original model allowing researchers to correct for any location and temporal errors compounded through earlier time-steps.

Chapter 8

8. General discussion

In this chapter, the major results from the previous chapters are discussed in relation to the aims and objectives of the thesis. Also, a number of recommendations are given for future research regarding species distribution and dispersal models. Finally, concluding remarks are given.

8.1 Uncertainty in species distribution modelling

The primary aim of this thesis was to develop methods that could address particular sources of uncertainty in correlative distribution model predictions. Seven specific objectives were studied under this aim. As a result, some new and some improved methods that reduce uncertainty in these models have been proposed in Chapters 3, 4 and 5.

8.1.1 Pseudo-absence data generation (Chapter 3)

One of the main components that introduces uncertainty to presence-absence correlative species distribution models is the method used (or lack thereof) to select pseudo-absences for SDMs (Lobo *et al.*, 2010). *The first objective* of this study (addressed in Chapter 3) was to evaluate the effect of pseudo-absence selection methods on individual model performance as well as model consensus. Accordingly, three widely used pseudo-absence selection methods and a new method developed in this study were compared. The results showed that the pseudo-absence selection method had a significant effect on the performance of individual models. Simple random selection of pseudo-absence points was heavily reliant on the prevalence of presence points and had an inconsistent effect on the models.

Therefore, simple random selection was not found to be a reliable pseudo-absence selection method, as discussed by Lobo *et al.* (2010). The second pseudo-absence selection method, which is based on simple random selection within a constrained geographical space, had a negative effect on both individual model performance and model consensus. Both the widely used two-step pseudo-absence selection method, which is based on environmental profiling followed by random sampling, and the three-step method developed in this study resulted in high individual model performance scores, however the latter resulted in improved consensus among model predictions.

The second objective of this study covered in Chapter 3 was to develop an improved pseudo-absence selection method that balances the geographical and environmental space when selecting pseudo-absence points. The proposed three-step pseudo-absence selection method resulted from an investigation of the most used pseudo-absence selection methods and the recommendations made from previous studies regarding the need for an improved method (Lobo *et al.*, 2010; Sinclair *et al.*, 2010). The methods reviewed in Chapter 3 either optimize pseudo-absence selection for the geographical space or the environmental space. Therefore, achieving a balance between environmental and geographical space was missing in existing methods. The two-step pseudo-absence selection method is the most used method and resulted in good model performance. However, because environmental profiling in this method is performed without geographical constraint, environmentally extreme pseudo-absences may be included in the pseudo-absences selected. Selection of extreme environmental pseudo-absences leads to an overly discriminated species distribution prediction where local variation is ignored. Moreover, models using highly discriminated presence/pseudo-absence points can easily overfit and therefore lose their generality (Lobo *et al.*, 2010).

The three-step method provides a way of balancing the information in geographic space with environmental space while selecting pseudo-absence points. This means the method allows ecologically meaningful boundaries to be set thus removing irrelevant geographical areas that are far distant from the occurrence points (VanDerWal *et al.*, 2009; Lobo *et al.*, 2010), thereby excluding extreme environmental points. The method still provides high

environmental discrimination between presence and pseudo-absence points based on robust classifiers as recommended by Chefaoui and Lobo (2008).

When compared with the two-step model, the three-step method resulted in comparable model performance. Additionally, and equally important, model predictions showed better consensus when trained with the three-step pseudo-absences.

8.1.2 Data, dimension reduction and model type (Chapter 4)

"Models are not like religion. You can have more than one… and you don't have to believe them"
Daniel Pauly and Villy Christensen

Discrepancy among model results is one of the factors behind the reluctance to use correlative species distribution models for species distribution projections for future climates (Buisson *et al.*, 2010; Guisan *et al.*, 2013). Kriticos *et al.* (2013) suggest that if models cannot agree when predicting for current conditions, then the confounding errors generated by climate projections will make future species distribution predictions even more unreliable. Clearly, there is a fundamental need to further our understanding of all factors that cause model discrepancy. While, the differences in the algorithms used for SDM's is an obvious reason for differences between predictions as shown in Chapter 4, the issue becomes further complicated when the predictions of different models vary with different modelling scenarios based on different species, variable type, and dimension reduction method.

Objective three of this study was aimed at studying the effect of using climatic and topographic variables additional to the usual temperature and precipitation derived variables for global species distribution studies. The analysis of the frequency of variables selected for the factorial study in Chapter 4 showed that variables from P2, the dataset with 35 variables, were selected as often as the variables from the P1 dataset that contains the 19 Bioclim variables usually used in global and regional SDM studies. Moreover, elevation from the predictor dataset P3 was also frequently selected for the different scenarios for the five species studied in that Chapter. Provided that a data-based variable selection method rather than expert knowledge is used, it is recommended to use BIOCLIM35 variables plus elevation for global species distribution studies.

The fourth objective of this study was to investigate the use of linear and non-linear dimension reduction methods on model performance of SDMs. In Chapter 4, it was shown that dimension reduction had a significant effect on model performance depending on the data and model used. The random forest variable selection method, where no feature construction was involved and raw variables were used for model training, gave a superior result compared with the principal component analysis (PCA) and non-linear principal component analysis (NLPCA) methods. Models trained on PCA transformed data had a generally low performance as reported by Dormann *et al.* (2008). However, the optimum model prediction for *D. v. virgifera* out of the different combinations in the factorial study was based on PCA transformed data. That prediction was able to identify the original native range of *D. v. virgifera* in Central America, which was not identified in the study in Chapter 3 as well as the studies carried out by Dupin *et al.* (2011) and Aragón *et al.* (2010), nor by using a fitted process-based modelling prediction (Kriticos *et al.*, 2012a). Such results show that, even though untransformed variables with a robust variable selection method like random forest, may give good model performance most of the time, but sometimes there may be cases where dimension reduction methods that involve feature construction (e.g. PCA) are needed to bring out the underlying pattern in the variables used to appropriately characterise the potential distribution of a species. Therefore, it is necessary to work with multiple scenarios of dimension reduction for species that are undergoing geographical and host range expansion as was the case with *D. v. virgifera* in Chapter 4.

To my knowledge, the factorial study in Chapter 4 is the first time the h-NLPCA has been used as a dimension reduction method for species distribution models. Unfortunately, the method resulted in over-prediction of most models trained on NLPCA transformed data. This was because the pseudo-absence points selected from the NLPCA data were highly localized/marginalized and discriminated on the environmental feature space which led to models over-fitting the pseudo-absence points. The models that over-fit the pseudo-absence points did not appropriately generalize the unsuitable habitat with extensive areas predicted as suitable for the target species. However, the good discriminatory power of the h-NLPCA is a quality that can be well utilized by SDMs that incorporate some regularization scheme

with model training to avoid over-fitting models (Dormann *et al.*, 2013). Additionally, I suggest that the h-NLPCA could probably be a robust method to perform feature selection for presence-only models like MAXENT due to the high discriminatory power obtained in results from Chapter 4. Alternatively, it could also be used with an OCSVM classifier to directly predict species distribution into a binary presence/absence output. This recommendation is made for presence-only models because these models predict species distribution based on information from occurrence data only and depend on a good characterization of the pattern of the occurrence data against the background data which they later classify into levels of suitability for the species. However further study is needed to confirm these assertions.

The fifth objective of this study was closely related to objectives three and four, as it was aimed at investigating the interactions between different modelling components on model performance. As shown in Chapter 4, variation in model performance was explained by differences in model type (MT), species data (SP), and dimension reduction methods (DR). Model type (MT) was found to be the major source of variation in line with previous studies by (Dormann *et al.*, 2008; Buisson *et al.*, 2010). In addition, the assessment of spatial standard errors of model predictions showed that, variation between model predictions according to model type (MT) was uniform across species, whereas the prediction variability according to other modelling components differed by species. This result supports the results of the multivariate analysis performed on model performance scores by confirming that model type has an important and consistent role in the accuracy of species' distribution predictions through a secondary analysis. Furthermore, the standard error of predictions according to model type show a distinct spatial variability that is independent of occurrence locations. This observation may explain why species' distribution predictions differ for the same species and location, as it shows that the capability of models to predict for areas away from occurrence points differ. However, in depth study that quantifies the level of spatial autocorrelation between occurrence points and spatial standard error patterns is needed to confirm this observation.

In Chapter 4, it was shown that it was possible to improve model performance by changing predictors or dimension reduction method within a multiple scenario modelling framework. This result shows, that even if model type is the major source of variation in model performance, by using the appropriate dimension reduction and predictor data (sets of climatic/environmental variables) it is possible to improve model performance. This has implication for model consensus, where using appropriate dimension reduction methods for the respective models in a multi-model setting can result in models giving comparable performance and prediction (Section 4.3.4).

To summarise, working within a multi-model and multi-scenario framework as suggested by Dormann *et al.* (2008) and Buisson *et al.* (2010) rather than performing a single model prediction will ensure that the most important modelling component, given the available data, is identified even in exceptional cases. The results from Chapter 4 also suggest that constituting species distribution models using some kind of framework that tests at least basic combinations of different predictor data and dimension reduction methods along with multiple models, is required to identify optimum conditions in which models perform best given the available species and environmental data.

The sixth objective of the study, also addressed in Chapter 4, was aimed at developing measures that could be used for model selection to complement confusion matrix based model performance measures. The analysis in Chapter 4 supported previous cautions by Allouche *et al.* (2006), Jiménez-Valverde *et al.* (2008) and Lobo *et al.* (2008) about complete reliance on confusion matrix based validation measures, such as Kappa and AUC to evaluate model performance. A major result from the research in Chapter 4 showed clearly that using the highest Kappa or AUC scores as a criterion for model selection is not always reliable (Section 4.3.7). Their unreliability is specifically because those models with highest Kappa or AUC may have over-fitted. There is no real way of knowing if a model has over-fitted other than a subjective assessment of model predictions. For this reason, cross-validation is performed to test both how well the model fit the data and to control for model over-fitting. However, in reality, test data often lack the power to identify over-fitting models because of low sample size, such that the test data may not be representative of the

whole environmental space in the case of large scale studies (global, regional). Moreover, even if models with the highest Kappa and AUC scores do not over-fit, it does not necessarily mean the highest score models are the only ones that can best predict the distribution of the target species. This result is shown in Chapter 4 where for all five species the models with maximum Kappa score did not have a significantly different Kappa from the next 3-5 high performing models, depending on the species. To overcome this problem, model cross-validation error was used to discriminate among models with similar Kappa for best model selection. It is worth noting that cross-validation error can be easily adapted for other multi-model species distribution prediction frameworks.

Clearly, measures of model performance like Kappa and AUC are important because they show how well a model fits the training data, and how well a model can generalize to new data. However, species distribution models trained and tested on limited data are often used to project species distribution over a much larger spatial extent and environmental domain. The relative cover indicators (RCIs) developed in Chapter 4 provide complementary validation methods to Kappa and AUC. The RCIs are independent of the models. The RCIs evaluate how well the data used for training and testing the models have covered the background data which the models are supposed to classify. The RCIs work for research designs where multiple scenarios of model components are involved in a modelling framework. This means, it will be possible to specify which dimension reduction method will work best based on the occurrence data and predictors used to provide an optimum medium for the models. This also means the RCIs can effectively evaluate modelling procedures before model training. Other studies have suggested using additional methods to confusion matrix based model performance measures. One study in particular by Engler *et al.* (2004) suggested using the minimum predicted area (MPA), to avoid models from over-predicting. The MPA is a similar method to the RCI approach in that it does not depend on how well the models trained on the sample data, rather it evaluates the final prediction of the models. Ideally one can use the RCIs as a pre-modelling test where the appropriate variables, and dimension reduction are chosen given the data available for the study. Then, use confusion matrix based methods to evaluate the model training performance, and finally

use a method like the MPA for post-modelling performance in which the prediction itself is assessed. Such holistic performance measures should give better confidence in correlative model results as the different levels of the modelling process could be evaluated using the set of performance measures mentioned above.

8.1.3 Multi-modality in occurrence data (Chapter 5)

The seventh objective of this thesis was to investigate the effect of variation within occurrence data on prediction accuracy and specify methods that enable improved species distribution prediction when using such occurrence data. One of the frustrations within prediction science is not being able to achieve the required accuracy with respect to predictions despite careful model specification practices. An area that could be investigated to address this problem is to examine the nature of the input data itself (presence data set) (Stockwell & Peterson, 2002; Dupin *et al.*, 2011; Chapter 5). In Chapter 5, I have shown using two case studies that accounting for the multi-modality of the ecological dataset, increases prediction accuracy.

The first case study was atypical in that it involved failure of all models to predict suitability of an area even when presence points were sampled from the same area. The original native range (in Central America) of *D. v. virgifera* was not predicted by the model that had the best Kappa, AUC as well as the other model performance methods presented in Chapter 3. The native Central American range of *D. v. virgifera* was also not predicted by two other studies using a similar modelling framework as that used in this research (Aragón *et al.*, 2010; Dupin *et al.*, 2011). Furthermore, another study using a fitted process-based prediction (CLIMEX) has also reported the Central American range as not suitable for *D. v. virgifera* (Kriticos *et al.*, 2012a).

In Chapter 4, the Central American range for *D. v. virgifera*, was predicted by the optimum model (SVM). The model in Chapter 4 was based on predictor data transformed using PCA as a dimension reduction method, which enabled the selected optimum model (SVM) to access better constructed feature data that can explain the variation within the *D. v. virgifera* presence data. However, most studies do not use any kind of dimension reduction method, therefore, why the native range (Central America) was not predicted using untransformed

variables was investigated in Chapter 5. Subsequently, basic statistical bimodality tests showed that the *D. v. virgifera* occurrence data had two distinct components (presences from the Central American range and presences from North America and Europe). I hypothesised that the variation between these sets of presences could be the reason why the original native range of *D. v. virgifera* was under-predicted in the study conducted in Chapter 3 as well as the other external studies mentioned above. By separately predicting the potential species distribution of *D. v. virgifera* according to presences from the two distinct components of the *D. v. virgifera* occurrence data and combining the predictions, it was possible to predict the Central American range of *D. v. virgifera.* This first case study of Chapter 5 showed that bimodality (multi-modality) in presence data could lead to under-prediction of the potential species distribution and that by using a combination of models to address the individual components within an occurrence data set, it is possible to better characterize the potential distribution of a species.

Identifying multi-modality is often very difficult and computationally intensive as the presence and/or absence of a species is affected by multiple factors that go beyond a few environmental variables used in the models. In other words, investigations for multi-modality might need to be based in a hyper-dimensional space instead of a low dimensional environmental space of few environmental variables. An alternative method, was demonstrated for cases where multi-modality might exist in the second case study on *Pieris brassicae* in Chapter 5. In that study, the investigation of specific biological traits that might create multi-modality in a presence dataset resulted in the identification of unique data components that needed to be modelled separately. Before *P. brassicae* was selected as the second case study of Chapter 5, a number of preliminary modelling exercises undertaken as a preparation for the study in Chapter 6, showed that all models failed to predict the newly invaded locality of *P. brassicae* in New Zealand. Further investigation about the species revealed that a particular population of *P. brassicae* in South Spain and Portugal undergoes summer diapause (aestivation) in addition to the winter diapause that is common to all *P. brassicae* populations.

The assumption was that, if a particular population of a species has a local adaptation, it might reflect on the geographical occurrences of those populations. Indeed, the aestivating population of *P. brassicae* is described as having a permanent geographical cline, where their population is limited to south of the Pyrenees (Spieth *et al.*, 2011), despite a recurrent immigration of populations coming from Central and Northern Europe. If there are such distinct presences in an occurrence data, the variables selected based on the complete presence points, might not necessarily reflect the variables that best explain the presence points recorded from the locally adapted populations. Based on the analysis in Chapter 5, the variables selected based on the complete set of presence points and the non-aestivating presences were similar, but different variables were selected for the aestivating presences. Such a result might be expected, because a small set of localized presences are expected to be explained by fewer variables. Therefore, a method used in geographical data analysis, namely, the directional standard deviational ellipse (Gong, 2002) was adapted to determine if the variation between the aestivating and non-aestivating populations of *P. brassicae* led to under-prediction of its potential distribution in the newly invaded locality in New Zealand. The results, showed that the variables selected based on all presence points, or the larger component (non-aestivating presences) did not explain the aestivating presences well, which led to masking of their contribution towards the potential global distribution of *P. brassicae*. Separately modelling these two different classes of presences of *P. brassicae*, led to correctly identifying the newly invaded range of *P. brassicae* in New Zealand.

The results of the research presented in Chapter 5 confirmed that modellers need to be alert to the possibility that biological traits that differ between populations of the same species may indicate different components in a presence dataset that should be modelled using mixed models for better prediction and increased accuracy. It is important to recognise, that such biological variation within the same species is likely to have some kind of associated geographical restriction that might affect species distribution models. Although, some biological variation can be independent of geographical variation, for example, genetic variation caused by change of hosts rather than geography as discussed by Phillips *et al.* (2008). In summary, it is recommended that the modeller needs to understand as much

about the biology of the species modelled as possible, so that any multi-modality in the presence dataset can be identified. This is an obvious requirement but has not been strongly suggested compared with the recommendation in many previous studies to use geographical variation to assess genetic variation and morphological differentiation (Escudero *et al.*, 2003; Dlugosch & Parker, 2008; Eckert *et al.*, 2008; Zapata & Jiménez, 2012).

8.2 Hybrid Correlative-Mechanistic Modelling (Chapter 6)

"If the only tool you have is a hammer, you tend to see every problem as a nail."

Abraham Maslow

The second aim and *eighth objective* of this thesis was to investigate the use of simple mechanistic models to reduce uncertainty about correlative model predictions by hybridising the results from the mechanistic and correlative modelling approaches.

Correlative models primarily depend on large samples of species occurrence data for sound species distribution predictions. As Diamond *et al.* (2012) noted, there comes a time when one should draw a line on whether to collect more data as opposed to using more complex models or improving the models being used. Essentially, there is a limit to the accuracy with which the unknown data can be inferred using correlative species distribution models.

However, decisions have to be made about the potential for species establishment in new areas. When there is inadequate occurrence data to represent the whole environmental range of a species the use of mechanistic models to complement correlative model predictions have been recommended repeatedly (Buckley *et al.*, 2010; Elith *et al.*, 2010; Dormann *et al.*, 2012- & references within).

While this might seem to be a straightforward recommendation, there is often a lack of physiological data and functional relationships of a species response to its environment and hence one of the main reasons for using correlative models. In Chapter 6 it was demonstrated that a minimally parameterized mechanistic model can be used to complement correlative models when data is sparse, by adapting the method developed by Monahan (2009) and Tingley *et al.* (2009). The generalized simple mechanistic model was parameterized using lethal thermal thresholds of the species studied (*P. brassicae*). Even if

this model was not complex, the simple mechanistic model was instrumental for defining the potential thermal niche of the species, which allowed over-prediction by the correlative species distribution model to be identified (Section 6.3.3.1).

The over-prediction by the correlative models was identified by simply overlaying the results from the simple mechanistic model and the correlative models. However, to fill in areas that were under-predicted by the correlative model with predictions from the mechanistic model, it was necessary to hybridize the outputs of the two approaches (Section 6.3.3.2). The hybrid suitability map, identified more suitable areas for *P. brassicae* in the North Island of New Zealand. In Section 6.3.2, it was shown that correlative model predictions improve as more presence points that better represent the environmental range of the species, are included. The newly identified suitable areas in the Northern Island of New Zealand and the correction for over-predicted areas globally, gives a good example of the use of hybrid predictions to complement correlative species distribution models in case of under- and over-prediction respectively.

8.3 Species-landscape interactions (Chapter 7)

The third aim of this thesis was to investigate multi-scale integration of global, regional and local data for characterization of the landscape on which dispersal studies are carried out. As part of this aim, the ninth objective was to specifically investigate if recoding of heterogeneous landscapes with varying degrees of composition across space could result in better dispersal rate and pattern prediction. The great white butterfly (*Pieris brassicae*) was used as a case study, and its spread, since it established in New Zealand in 2010 was modelled using an individual based spatially explicit population model (MDiG) (Pitt *et al.*, 2009).

The determination of an optimum landscape resolution, configuration and composition is a subject of ongoing discussion, so it was decided to use a simple approach comprising selective recoding of the landscape for areas that need more detailed configuration (Chapter 7, Section 7.3.5.4). *P. brassicae*, was introduced in an urban area (Nelson), and has spread throughout home gardens and green spaces within the city (Phillips *et al.*, 2013). However,

in the land use data available for New Zealand, all landscape details in urban areas are currently generalised into one class. A high resolution satellite image was used to recode the generalized urban areas for a better representation of the urban landscape in terms of the survival probability of the *P. brassicae*. Dispersal that was modelled on the recoded landscape resulted in a slower dispersal rate compared with the generalized landscape. It was also shown that the dispersal model had higher specificity on the recoded landscape than the landscape based on the generalized land use data. That result is important because achieving high specificity in species dispersal models is crucial and has direct implications in decision making (Pitt *et al.*, 2011). For example, dispersal patterns obtained from a low specificity model can give an unwarranted pessimistic view in terms of the cost of eradicating an invasive species that could lead to eradication attempts being abandoned. The longer an invasive species stays in its newly invaded habitat the more likely they re-structure interactions in the eco-system to the point they may replace functional roles of native species (Beggs & Rees, 1999; Gardner-Gee & Beggs, 2013). Sometimes when species reach such a stage any eradication effort may potentially be more of a disadvantage to the ecosystem (Zavaleta *et al.*, 2001).

Selective recoding of a landscape can also be used to better utilise available land use data that may not be up to the standard for a given dispersal study. Most national data sources used for high resolution spatial modelling such as vegetation indices, land cover, and soil maps are scaled to match the national spatial data framework. Even though the data may have originated from high resolution data sources, if they are generalized to match the common scale used in the country, it will not be possible to access the required level of detail for the specific species being modelled. For such cases, recoding certain landuse types that have strong species-landscape interaction can be a cost-effective way of utilizing available landscape data for species dispersal modelling.

In Chapter 7, data from multiple scales were used to produce the landscape on which the *P. brassicae* dispersal was carried out. These layers were, the hybrid climatic suitability layer (global scale), the land cover and growing degree data (regional scale) and the high resolution satellite imagery (local scale). Such integration of various information about the

landscape is useful for characterising a particular landscape. To ensure the most detailed dataset was not influenced by the global scale inputs, a weighted scoring scheme (Section 7.2.5.5) was employed. That scheme could be used in similar studies to easily incorporate data from different scales into the landscape. Perhaps an even better outcome could be obtained by multiplying these different layers instead of using the additive scheme shown in this study, if the explicit effect of the different layers on species' survival probability is known (Kean & Phillips, 2013a).

8.4. Future research

8.4.1 Pseudo-absence generation

The procedure used to determine the geographical boundary for the pseudo-absence selection method developed in Chapter 3, gives a better method for identifying ecologically meaningful geographical boundary than any currently used methods.

However, the procedure could be made more objective if a change detection algorithm that works on continuous data is employed to demark the appropriate boundary distance. In its current form, discrete distance intervals are investigated to detect a change in variable contribution.

The merits and demerits of true absences for species distribution modelling have been discussed in previous studies (Stockwell & Peters, 1999; Wisz & Guisan, 2009). A study by Hanberry *et al.* (2012) reported that model performance increased with the additional use of true (surveyed) absences along with environmentally profiled pseudo-absence points. Indeed, since their model was based on tree species their confidence in the surveyed absence points can be very high. However, the use of true absence points cannot readily generalize to all species distribution models, especially to those constructed for mobile species. Some insect species are too small and probably cryptic, therefore limiting the accuracy of surveyed true absence points (Hirzel *et al.*, 2002). If high confidence in true absence points could be achieved by repeated surveys or other monitoring techniques, then applying Hanberry *et al.* (2012)'s proposed method with the thee-step pseudo-absence method presented in Chapter 3 might also have a positive outcome of giving more consensus among model predictions.

8.4.2 Research design for species distribution models

A number of studies including this one have established that variation in model results for the same species depend on the data, the model specification and model type (Elith *et al.*, 2006; Pearson *et al.*, 2006; Dormann *et al.*, 2008; Elith & Graham, 2009; Buisson *et al.*, 2010; Senay *et al.*, 2013). Additionally, in this study it was shown that some models could perform better with specific environmental variables, pre-processing methods or a unique combination of both (Chapter 4 & Chapter 5). Therefore, the development of modelling frameworks that not only facilitate the use of multiple models, but also allow for multiple data pre-processing techniques is recommended. More important, only a few of the available data mining and pre-processing methods are currently being used for species distribution modelling. As the number of variables that become available for correlative studies increase, it will be important to adapt powerful data pre-processing methods to identify meaningful data for models, for a robust species distribution prediction.

8.4.3 Hybrid modelling

There are a number of opportunities to use hybrid correlative-mechanistic models as well as other combinations of hybrid models to improve accuracy of species distribution predictions. Similarly, such opportunity means greater challenges as with all new approaches there are gaps in our knowledge required to obtain robust predictions from environmental or physiological data.

An area that requires more research is to define generalized rules where physiological suitability obtained from mechanistic models can be made comparable to suitability values obtained from correlative models. Such rules will promote robust methods for hybridizing results of different models. This is a difficult challenge as the assumptions on which mechanistic and correlative models are based are completely different. However, standardized results from such models should be possible, if an output rather than process orientated method is used. An output orientated hybridizing method approach develops frameworks that standardize the different outputs of these models in terms of suitability for the species instead of standardizing the procedure by which the different modelling approaches predict species distributions. For this to work, the individual models have to be

individually validated and their predictions accepted when it is within a pre-determined level of uncertainty. Once the output of the individual models are already rated in terms of suitability to the species, it will be easier to work out a criteria based on which, results from correlative and mechanistic models could be hybridized.

If there is a spatial uncertainty layer for both the correlative and mechanistic models, for instance, results could be combined based on the magnitude of the positional uncertainty, leading to a hybridized prediction that has less spatial uncertainty compared with the input correlative and mechanistic predictions.

8.4.4 Species dispersal modelling

Individual based models are increasingly becoming the model of choice for species dispersal simulations (Chapter 7; Grimm, 1999; Pitt, 2008 - & references within; Savage & Renton, 2014). In the ecological implementation of individual based models, the spatial position of the individual at each time step is as important as the final stable pattern (equilibrium).

Equilibrium in reality is an elusive concept, we almost never see a species in equilibrium with its niche because of complex interactions within the landscape. For example, availability of host species, biotic competition, human impact etc. Determining an equilibrium state for a species is even more difficult when modelling is undertaken for non-pristine ecosystems where continued disturbance changes key parameters (Boyd, 2012).

However, equilibrium is adequately defined in classical ecology (UchmaDski & Grimm, 1996) and while the understanding is there, we just lack the actual models and associated assumptions that could prescribe the optimum level of landscape heterogeneity, level of complexity of species-landscape and species-species interactions to accurately predict when a species could be at an equilibrium with its niche at a future point in time. The capability to accurately estimate the spatial pattern of an alien species invasion at any time in the future including the time it takes to reach equilibrium within its new habitat is paramount for planning eradication strategies or to even decide whether any eradication effort is necessary or possible.

While selective re-coding of certain areas of the landscape based on species attributes as demonstrated in Chapter 7 could greatly benefit studies case by case, a generalized and exhaustive study that could elucidate a possible relationship between species attributes and mode of dispersal with optimum landscape resolution, configuration and composition is greatly needed.

8.5 Concluding remarks

"Errors using inadequate data are much less than those using no data at all."

Charles Babbage

Finally, if model discrepancy is taken as a measure of uncertainty in correlative model predictions (Kriticos *et al.*, 2013), then their consensus (Chapter 3 & 4) within reason could indicate improved accuracy within a modelling framework (within reason because there is a chance models could wrongly agree). Additionally, hybrid modelling that focuses on maximising the advantages from correlative and mechanistic modelling while minimizing the compounded error from the individual limitations of the models (as in Chapter 6) can be one way of increasing confidence in predictions of current species distribution. Modelling frameworks should at least be valid for the current situation in order to discuss their use for future species distribution predictions. I believe the improvements recommended in this thesis as well as other recent studies involving species distribution models have merits refining such models for use in future climate predictions (Sinclair *et al.*, 2010).

Some researchers have consistently suggested that different ecological models fit into a continuum or a gradient based on their different attributes, instead of being simply dichotomized into classes of correlative and mechanistic models (Hirzel & Le Lay, 2008; Jiménez-Valverde *et al.*, 2008; Sillero, 2011; Dormann *et al.*, 2012). Such definition has a practical implication for improving current and future modelling attempts. First, modellers can discard the idea of the need to specialize in any of the classified modelling camps. Second, while the different species distribution model approaches are based on different assumptions, by hybridizing these different models, it is possible to model data considered inadequate according to one type of modelling approach.

Finally, it is imperative that ecological models whether they are static, like the species distribution models discussed in Chapter 3-6 or dynamic dispersal models as demonstrated in Chapter 7 are specified in a such a way that the associated uncertainty associated with model results is clearly presented so that end-users could make informed decisions about the potential for invasive species establishment.

9. References

Abate, T., Ebro, A., & Nigatu, L. (2009). Pastoralists perceptions and rangeland evaluation for livestock production in South Eastern Ethiopia. *Livestock Research for Rural Development, 21*

Abbott, K. L. (2005). Supercolonies of the invasive yellow crazy ant, *Anoplolepis gracilipes*, on an oceanic island: Forager activity patterns, density and biomass. *Insectes Sociaux, 52*(3), 266-273. doi:10.1007/s00040-005-0800-6

Abbott, K. L., & Green, P. T. (2007). Collapse of an ant–scale mutualism in a rainforest on Christmas Island. *Oikos, 116*(7), 1238-1246. doi:10.1111/j.0030-1299.2007.15629.x

Abbott, K. L., Green, P. T., & O'Dowd, D. J. (2014). Seasonal shifts in macronutrient preferences in supercolonies of the invasive Yellow Crazy Ant *Anoplolepis gracilipes* (Smith, 1857)(Hymenoptera: Formicidae) on Christmas Island, Indian Ocean. *Austral Entomology,*

Abbott, K. L., Harris, R., & Lester, P. (2005). *Invasive ant risk assessment: Anoplolepis gracilipes* (No. Landcare Research contract report for Biosecurity New Zealand). Wellington, New Zealand.

Acosta, C. A. (2002). Spatially explicit dispersal dynamics and equilibrium population sizes in marine harvest refuges. *ICES Journal of Marine Science: Journal du Conseil, 59*(3), 458-468. doi:10.1006/jmsc.2002.1196

Admasu, D. (2008). *Invasive plants and food security: the case of Prosopis julifora in the Afar region of Ethiopia*. Paper presented at the IUCN Conference.

Aguirre-Gutiérrez, J., Carvalheiro, L. G., Polce, C., van Loon, E. E., Raes, N., Reemer, M., & Biesmeijer, J. C. (2013). Fit-for-Purpose: Species Distribution Model Performance Depends on Evaluation Criteria – Dutch Hoverflies as a Case Study. *PLoS ONE, 8*(5), e63708. doi:10.1371/journal.pone.0063708

Ahmed, S., Goria, M., & Hussein, A. (2008). Gamma mixture: bimodality, inflexions and L-moments. *Communications in Statistics—Theory and Methods, 37*(8), 1147-1161.

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of applied ecology, 43*(6), 1223-1232.

Altieri, A. H., van Wesenbeeck, B. K., Bertness, M. D., & Silliman, B. R. (2010). Facilitation cascade drives positive relationship between native biodiversity and invasion success. *Ecology, 91*(5), 1269-1275. doi:10.1890/09-1301.1

Amezaga, I. (1997). Forest characteristics affecting the rate of shoot pruning by the pine shoot beetle (Tomicuspiniperda L.) in Pinus radiata D. Don and P. sylvestris L. plantations. *Forestry, 70*(2), 129-137.

Andersen, M. C., Adams, H., Hope, B., & Powell, M. (2004). Risk analysis an official publication of the Society for Risk Analysis *Risk analysis an official publication of the Society for Risk Analysis, 24*(4), 893-900.

Anderson, D. R. (2007). *Model Based Inference in the Life Sciences: A Primer on Evidence*: Springer.

Anderson, L. (2005). California's Reaction to *Caulerpa taxifolia*: A Model for Invasive Species Rapid Response. *Biological Invasions, 7*(6), 1003-1016. doi:10.1007/s10530-004-3123-z

Andow, D. A., Kareiva, P. M., Levin, S., & Okubo, A. (1990). Spread of invading organisms. *Landscape Ecology, 4*(2-3), 177-188. doi:10.1007/BF00132860

Andrew, M. E., & Ustin, S. L. (2009). Habitat Suitability modelling of an invasive plant with advanced remote sensing data. *Diversity and Distributions, 15*, 628. doi:10.1111/j.1472-4642.2009.00568.x

Angassa, A., & Oba, G. (2008a). Effects of management and time on mechanisms of bush encroachment in southern Ethiopia. *African Journal of Ecology, 46*(2), 186-196. doi:10.1111/j.1365-2028.2007.00832.x

Angassa, A., & Oba, G. (2008b). Herder Perceptions on Impacts of Range Enclosures, Crop Farming, Fire Ban and Bush Encroachment on the Rangelands of Borana, Southern Ethiopia. *Human Ecology, 36*(2), 201-215. doi:10.1007/s10745-007-9156-z

Ansari, M., Hasan, F., & Ahmad, N. (2012). Influence of various host plants on the consumption and utilization of food by Pieris brassicae (Linn.).

Aragón, P., Baselga, A., & Lobo, J. M. (2010). Global estimation of invasion risk zones for the western corn rootworm *Diabrotica virgifera virgifera*: integrating distribution models and physiological thresholds to assess climatic favourability. *Journal of Applied Ecology, 47*(5), 1026-1035. doi:10.1111/j.1365-2664.2010.01847.x

Araújo, M. B., & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography, 33*(10), 1677-1688. doi:10.1111/j.1365-2699.2006.01584.x

Araújo, M. B., & New, M. (2007). Ensemble forecasting of species distributions. *Trends in Ecology & Evolution, 22*(1), 42-47.

Araújo, M. B., & Pearson, R. G. (2005). Equilibrium of species' distributions with climate. *Ecography, 28*(5), 693-695. doi:10.1111/j.2005.0906-7590.04253.x

Araújo, M. B., Pearson, R. G., Thuiller, W., & Erhard, M. (2005). Validation of species–climate impact models under climate change. *Global Change Biology, 11*(9), 1504-1513.

Araújo, M. B., & Peterson, A. T. (2012). Uses and misuses of bioclimatic envelope modelling. *Ecology, 93*(7), 1527-1539. doi:10.1890/11-1930.1

Ashman, K. M., Bird, C. M., & Zepf, S. E. (1994). Detecting bimodality in astronomical datasets. *Astron. J., 108*, 2348-2361. doi:10.1086/117248

Astrom, M., Dynesius, M., Hylander, K., & Nilsson, C. (2007). Slope aspect modifies community response to clear-cutting in boreal forests. *Ecology, 88*(3), 749.

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

Austin, M., & Van Niel, K. P. (2011). Improving species distribution models for climate change studies: variable selection and scale. *Journal of Biogeography, 38*(1), 1-8. doi:10.1111/j.1365-2699.2010.02416.x

Avtzis, D., & Avtzis, N. (1999). Control of the most dangerous insects of Greek forests and plantations. *Proceedings: Integrated management and dynamics of forest defoliating insects*, 15-19.

Aydin, K. Y., McFarlane, G. A., King, J. R., Megrey, B. A., & Myers, K. W. (2005). Linking oceanic food webs to coastal production and growth rates of Pacific salmon *Oncorhynchus* spp. using models on three scales. *Deep Sea Research Part II: Topical Studies in Oceanography, 52*(5), 757-780.

Baker, M., Nur, N., & Geupel, G. R. (1995). Correcting biased estimates of dispersal and survival due to limited study area: theory and an application using wrentits. *The Condor, 97*(3), 663-674. doi:10.2307/1369175
Baker, W. L. (1993). Spatially Heterogeneous Multi-Scale Response of Landscapes to Fire Suppression. *Oikos, 66*(1), 66-71. doi:10.2307/3545196
Baldi, P., & Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks, 2*(1), 53-58. doi:10.1016/0893-6080(89)90014-2
Barber, A., Pellow, G., & Barber, M. (2011). *Carbon Footprint of New Zealand Arable Production – Wheat, Maize Silage, Maize Grain and Ryegrass Seed*: Ministry of Agriculture and Forestry. . (MAF Technical Paper No: 2011/97). Retrieved from http://www.fedfarm.org.nz/Files/2011-MPIGrainCarbon.pdf Accessed 30 Oct 2012
Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution, 3*(2), 327-338. doi:10.1111/j.2041-210X.2011.00172.x
Barlow, N. D., & Dixon, A. F. G. (1980). *Simulation of Lime Aphid Population Dynamics*: Centre for Agricultural Publishing and Documentation.
Barney, J. (2006a). North American History of Two Invasive Plant Species: Phytogeographic Distribution, Dispersal Vectors, and Multiple Introductions. *Biological Invasions, 8*(4), 703-717. doi:10.1007/s10530-005-3174-9
Barney, J. N. (2006b). A North American history of two invasive plant species: phytogeographic distribution, dispersal vectors and multiple introductions. *Biological Invasions, 8*, 706-715.
Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., Soberón, J., & Villalobos, F. (2011). The crucial role of the accessible area in ecological niche modelling and species distribution modelling. *Ecological Modelling, 222*(11), 1810-1819. doi:10.1016/j.ecolmodel.2011.02.011
Battisti, A. (1988). Host-plant relationships and population dynamics of the Pine Processionary Caterpillar *Thaumetopoea pityocampa* (Denis & Schiffermuller). *Journal of Applied Entomology, 105*(1-5), 393-402. doi:10.1111/j.1439-0418.1988.tb00202.x
Battisti, A., Longo, S., Tiberi, R., & Triggiani, O. (1998). Results and perspectives in the use ofBacillus thuringiensis Berl. var. kurstaki and other pathogens against*Thaumetopoea pityocampa* (Den. et Schiff.) in Italy (Lep., Thaumetopoeidae). *Anzeiger für Schädlingskunde, Pflanzenschutz, Umweltschutz, 71*(4), 72-76.
Battisti, A., Stastny, M., Buffo, E., & Larsson, S. (2006). A rapid altitudinal range expansion in the pine processionary moth produced by the 2003 climatic anomaly. *Global Change Biology, 12*(4), 662-671.
Battisti, A., Stastny, M., Netherer, S., Robinet, C., Schopf, A., Roques, A., & Larsson, S. (2005). Expansion of geographic range in Pine processionary moth caused by increased winter temperatures. *Ecological Applications, 15*(6), 2084-2096.

Beaumont, L. J., Gallagher, R. V., Thuiller, W., Downey, P. O., Leishman, M. R., & Hughes, L. (2009). Different climatic envelopes among invasive populations may lead to underestimations of current and future biological invasions. *Diversity and Distributions, 15*(3), 409-420. doi:10.1111/j.1472-4642.2008.00547.x
Beaumont, L. J., Hughes, L., & Poulsen, M. (2005). Predicting species distributions: use of climatic parameters in BIOCLIM and its impact on predictions of species' current and future distributions. *Ecological Modelling, 186*(2), 251-270. doi:10.1016/j.ecolmodel.2005.01.030
Beest, M., Elschot, K., Olff, H., & Etienne, R. S. (2013). Invasion Success in a Marginal Habitat: An Experimental Test of Competitive Ability and Drought Tolerance in *Chromolaena odorata*. *PLoS ONE, 8*(8), e68274. doi:10.1371/journal.pone.0068274
Beggs, J., & Rees, J. S. (1999). Restructuring of Lepidoptera communities by introduced *Vespula* wasps in a New Zealand beech forest. *Oecologia, 119*(4), 565-571. doi:10.1007/s004420050820
Beggs, J., Rees, J. S., & Harris, R. J. (2002). No evidence for establishment of the wasp parasitoid, *Sphecophaga vesparum burra* (Cresson) (Hymenoptera: Ichneumonidae) at two sites in New Zealand. *New Zealand Journal of Zoology, 29*(3), 205-211. doi:10.1080/03014223.2002.9518304
Beggs, J., Toft, R., Malham, J., Rees, J., Tilley, J., Moller, H., & Alspach, P. (1998). The difficulty of reducing introduced wasp (*Vespula vulgaris*) populations for conservation gains. *New Zealand Journal of Ecology, 22*(1), 55-63.
Begon, M., Townsend, C. R., & Harper, J. L. (2006). *Ecology: From Individuals to Ecosystems* (4 ed.). Oxford, United Kingdom: Blackwell Publishing.
BenDor, T. K., Metcalf, S. S., Fontenot, L. E., sangunett, B., & Hannon, B. (2006). Modeling the spread of the Emerald Ash Borer. *Ecological Modelling 197*, 221,224. doi:10.1016/j.ecolmodel.2006.03.003
Bilton, D. T., Freeland, J. R., & Okamura, B. (2001). Dispersal in freshwater invertebrates. *Annual Review of Ecology, Evolution, and Systematics, 32*, 160-165.
Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*: Oxford university press.
Boland, C., Smith, M., Maple, D., Tiernan, B., Barr, R., Reeves, R., & Napier, F. (2011). Heli-baiting using low concentration fipronil to control invasive yellow crazy ant supercolonies on Christmas Island, Indian Ocean. *Island invasives: eradication and management. Gland: International Union for Conservation of Nature*, 152-156.
Bonnemaison, L. (1965). Insect pests of crucifers and their control. *Annual Review of Entomology, 10*(1), 233-256.
Boyd, I. L. (2012). The art of ecological modelling. *science, 337*(6092), 306-307.
Bradley, B. A., & Mustard, J. F. (2006). Characterizing the landscape dynamics of an invasive plant and risk of invasion using remote sensing. *Ecological Applications, 16*(3), 1132.
Branson, T. F., & Krysan, J. L. (1981). Feeding and oviposition behavior and life cycle strategies of *Diabrotica*: An evolutionary view with implications for pest management. *Environmental Entomology, 10*(6)
Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32. doi:10.1023/a:1010933404324
Brockerhoff, E. G., Jactel, H., Goldarazena, A., Berndt, L., & Bain, J. (2006). *Risk assessment of European pests of Pinus radiata* (No. CLIENT REPORT No: 12216). New Zealand: New Zealand Forest Health Research Collaborative.
Brotons, L., Thuiller, W., Araújo, M. B., & Hirzel, A. H. (2004). Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography, 27*(4), 437-448. doi:10.1111/j.0906-7590.2004.03764.x
Brown, J. H., Whitham, T. G., Ernest, S. M., & Gehring, C. A. (2001). Complex species interactions and the dynamics of ecological systems: long-term experiments. *Science*, 643-650.
Buchsbaum, R., & Buchsbaum, M. (1957). *Basic Ecology*. Pittsburgh: The Boxwood Press.

Buckley, L. B., Urban, M. C., Angilletta, M. J., Crozier, L. G., Rissler, L. J., & Sears, M. W. (2010). Can mechanism inform species' distribution models? *Ecology Letters, 13*(8), 1041-1054. doi:10.1111/j.1461-0248.2010.01479.x

Buisson, L., Thuiller, W., Casajus, N., Lek, S., & Grenouillet, G. (2010). Uncertainty in ensemble forecasting of species distribution. *Global Change Biology, 16*(4), 1145-1157. doi:10.1111/j.1365-2486.2009.02000.x

Burgman, M. A., & Fox, J. C. (2003). Bias in species range estimates from minimum convex polygons: implications for conservation and options for improved planning. *Animal Conservation, 6*(1), 19-28. doi:10.1017/s1367943003003044

Burrough, P. A., & McDonell, R. A. (1998). *Principles of Geographical Information Systems*. New York: Oxford University Press.

Busby, J. R. (1986). A biogeoclimatic analysis of *Nothofagus cunninghamii* (Hook.) Oerst. in southeastern Australia. *Australian Journal of Ecology, 11*(1), 1-7. doi:10.1111/j.1442-9993.1986.tb00912.x

Busby, J. R., McMahon, J. P., Hutchinson, M. F., Nix, H. A., & Ord, K. D. (1991). BIOCLIM - a bioclimate analysis and prediction system. *Plant protection Quarterly, 6*(1), 8-9.

Cain, M. L., Milligan, B. G., & Strand, A. E. (2000). Long-distance seed dispersal in plant populations. *American Journal of Botany, 87*(9), 1217-1227.

Calabrese, J. M., Certain, G., Kraan, C., & Dormann, C. F. (2014). Stacking species distribution models and adjusting bias by linking them to macroecological models. *Global Ecology and Biogeography, 23*(1), 99-112. doi:10.1111/geb.12102

Caminade, C., Medlock, J. M., Ducheyne, E., McIntyre, K. M., Leach, S., Baylis, M., & Morse, A. P. (2012). Suitability of European climate for the Asian tiger mosquito Aedes albopictus: recent trends and future scenarios. *Journal of The Royal Society Interface, 9*(75), 2708-2717.

Carpenter, G., Gillison, A. N., & inter, J. W. (1993). DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation, 2*, 667-680.

Carpenter, J. M., & Glare, T. R. (2010). Misidentification of *Vespula alascensis* as *V. vulgaris* in North America (Hymenoptera: Vespidae; Vespinae). *American Museum Novitates*, 1-7. doi:10.1206/706.1

Carvalho, D. O., Costa-da-Silva, A. L., Lees, R. S., & Capurro, M. L. (2013). Two step male release strategy using transgenic mosquito lines to control transmission of vector-borne diseases. *Acta Tropica,*

Cavanaugh, K. C., Siegel, D. A., Raimondi, P. T., & Alberto, F. (2014). Patch definition in metapopulation analysis: a graph theory approach to solve the mega-patch problem. *Ecology, 95*(2), 316-328. doi:10.1890/13-0221.1

Chave, J., & Levin, S. (2004). Scale and Scaling in Ecological and Economic Systems. In P. Dasgupta & K.-G. Mäler (Eds.), *The Economics of Non-Convex Ecosystems* (Vol. 4, pp. 29-59): Springer Netherlands. Retrieved from 10.1007/1-4020-2515-7_2. doi:10.1007/1-4020-2515-7_2

Chefaoui, R. M., & Lobo, J. M. (2008). Assessing the effects of Pseudo-absence on predictive distribution model performance. *Ecological Modelling, 210*, 478-486. doi:10.1016/j.ecolmodel.2007.08.010

Chen, J., & Li, P. (2009). Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics, 37*(5A), 2523-2542.

Chevan, A., & Sutherland, M. (1991). Hierarchical partitioning. *The American Statistician, 45*(2), 90-96.

Chew, F. S., & Renwick, J. A. A. (1995). Host Plant Choice in *Pieris* Butterflies. In *Chemical Ecology of Insects 2* (pp. 214-238): Springer US. doi:10.1007/978-1-4615-1765-8_6

Chong, K.-F., & Lee, C.-Y. (2010). Inter- and Intraspecific Aggression in the Invasive Longlegged Ant (Hymenoptera: Formicidae). *Journal of Economic Entomology, 103*(5), 1775-1783. doi:10.1603/EC09256

Christensen, V., & Walters, C. J. (2004). Ecopath with Ecosim: methods, capabilities and limitations. *Ecological modelling, 172*(2), 109-139.

Chuine, I., & Beaubien, E. G. (2001). Phenology is a major determinant of tree species range. *Ecology Letters, 4*(5), 500-510. doi:10.1046/j.1461-0248.2001.00261.x

Ciosi, M., Miller, N. J., Kim, K. S., Giordano, R., Estoups, A., & Guillemaud, T. (2008). Invasion of Europe by the western corn rootworm, *Diabrotica virgifera virgifera*: multiple transatlantic introductions with various reductions of genetic diversity. *Molecular Ecology, 17*, 3622-3625. doi:10.1111/j.1365-294X.2008.03866.x

Clapperton, B. K., Tilley, J. A. V., Beggs, J. R., & Moller, H. (1994). Changes in the distribution and proportions of *Vespula vulgaris* (L.) and *Vespula germanica* (Fab.) (Hymenoptera: Vespidae) between 1987 and 1990 in New Zealand. *New Zealand Journal of Zoology, 21*(3), 295-303. doi:10.1080/03014223.1994.9517998

Coats, S. A., Tollefson, J. J., & Mutchmor, J. A. (1986). Study of migratory flight in the Western Corn Rootworm (Coleoptera: Chrysomelidae). *Environmental Entomology, 15*

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, 20*(1), 37-46. doi:10.1177/001316446002000104

Colautti, R. I., & MacIsaac, H. J. (2004). A neutral terminology to define 'invasive' species. *Diversity and Distributions, 10*(2), 135-141. doi:10.1111/j.1366-9516.2004.00061.x

Colwell, R. K., & Rangel, T. F. (2009). Hutchinson's duality: the once and future niche. *Proceedings of the National Academy of Sciences, 106*(Supplement 2), 19651-19658.

Conn, A. R., Gould, N. I. M., & Toint, P. L. (2000). *Trust Region Methods*: Society for Industrial and Applied Mathematics.

Cousens, R. D., & Cousens, J. M. (2011). Invasion of the New Zealand Coastline by European Sea-Rocket (*Cakile maritima*) and American Sea-Rocket (*Cakile edentula*). *Invasive Plant Science and Management, 4*(2), 260-263. doi:10.1614/IPSM-D-10-00060.1

Crepet, A., Paugam-Moisy, H., Reynaud, E., & Puzenat, D. (2000). *A modular neural model for binding several modalities.* Paper presented at the IC-AI

Crooks, J. A. (2005). Lag times and exotic species: the ecology and management of biological invasions in slow-motion. *Ecoscience, 12*(3), 316-329.

Csurhes, S., & Hankamer, C. (2012). *Pest animal risk assessment: Yellow Crazy ant (Anoplolepis gracilipes)* (No. CS 0975 02/12). Queensland, Australia: Biosecurity Queensland, Queensland Government.

Cunningham, E. P., Abusowa, M., Lindquist, D. A., E., S. A., & Vargas-Terán, M. (1992). The North Africa screwworm eradication programme. *Invertebrate Reproduction and Development 22* (1-3), 103-108.

D'Adamo, P., Sackmann, P., Corley, J. C., & Rabinovich, M. (2002). The potential distribution of German wasps (*Vespula germanica*) in Argentina. *New Zealand Journal of Zoology, 29*(2), 79-85. doi:10.1080/03014223.2002.9518292
Daves, C., Higgins, R., Sloderbeck, P., Wilde, G., Whitworth, R., Zhu, K., & Buschman, L. (2007). How Kansas crop consultants scout for western corn rootworms (Coleoptera: Chrysomelidae) in field corn. *AMERICAN ENTOMOLOGIST-LANHAM-, 53*(1), 8.
David, W. A. L., & Gardiner, B. O. C. (1962). Oviposition and the hatching of the eggs of *Pieris brassicae* (L.) in a laboratory culture. *Bulletin of Entomological Research, 53*(01), 91-109. doi:doi:10.1017/S0007485300047982
Davies, C. R., & Gilbert, N. (1985). A comparative study of the egg-laying behaviour and larval development of *Pieris rapae* L. and *P. brassicae* L. on the same host plants. *Oecologia, 67*(2), 278-281. doi:10.1007/BF00384299
Davis, M. A., Thompson, K., & Grime, J. P. (2001). Charles S. Elton and the dissociation of invasion ecology from the rest of ecology. *Diversity & Distributions, 7*(1-2), 97-102. doi:10.1046/j.1472-4642.2001.00099.x
Debarma, M., & Firake, D. M. (2013). Host generated cues alter the foraging behavior of Cabbage butterfly, *Pieris brassicae* and its larval parasitoids, *Cotesia glomerata* and *Hyposoter ebeninus*. *2013, 45*(2). P. brassicae, H. ebeninus, C. glomerata. doi:10.4081/jear.2013.e15e15
Derraik, J. G. B. (2004). Exotic mosquitoes in New Zealand: a review of species intercepted, their pathways and ports of entry. *Australian and New Zealand Journal of Public Health, 28*(5), 433-444. doi:10.1111/j.1467-842X.2004.tb00025.x
Devkota, B., & Schmidt, G. H. (1990). Larval development of *Thaumetopoea pityocampa* (Den. & Schiff.) (Lep., Thaumetopoeidae) from Greece as influenced by different host plants under laboratory conditions. *Journal of Applied Entomology, 109*(1-5), 321-330. doi:10.1111/j.1439-0418.1990.tb00059.x
Diamond, S. E., Nichols, L. M., McCoy, N., Hirsch, C., Pelini, S. L., Sanders, N. J., Ellison, A. M., Gotelli, N. J., & Dunn, R. R. (2012). A physiological trait-based approach to predicting the responses of species to experimental climate warming. *Ecology, 93*(11), 2305-2312. doi:10.1890/11-2296.1
Diaz-Uriarte, R. (2009). varSelRF: Variable selection using random forests (Version R package version 0.7-1.): http://ligarto.org/rdiaz/Software/Software.html Accessed 2012 Jun 13. Retrieved October 19, 2012. Available from http://cran.r-project.org/web/packages/varSelRF/index.html
Dillen, K., Mitchell, P. D., & Tollens, E. (2009). On the competitiveness of *Diabrotica virgifera virgifera* damage abatement strategies in Hungary: a bio-economic approach. *J. Appl. Entomol, 134*(5). doi:j.1439-0418.2009.01454.x
Diniz-Filho, J. A. F., Mauricio Bini, L., Fernando Rangel, T., Loyola, R. D., Hof, C., Nogués-Bravo, D., & Araújo, M. B. (2009). Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. *Ecography, 32*(6), 897-906. doi:10.1111/j.1600-0587.2009.06196.x
Dlugosch, K. M., & Parker, I. M. (2008). Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. *Molecular Ecology, 17*(1), 431-449. doi:10.1111/j.1365-294X.2007.03538.x
DOC. (2013a). *Great White Butterfly News*. Nelson, New Zealand.
DOC. (2013b). Great white butterfly: Help stop this major new pest. In D. o. Conservation (Ed.) (pp. 2) http://www.doc.govt.nz/Documents/conservation/threats-and-impacts/animal-pests/nelson-marlborough/great-white-butterfly-factsheet.pdf
Donovan, B. J. (1984). Occurrence of the common wasp, *Vespula vulgaris* (L.) (Hymenoptera: Vespidae) in New Zealand. *New Zealand Journal of Zoology, 11*(4), 417-427. doi:10.1080/03014223.1984.10428256
Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., & Leitão, P. J. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography, 36*(1), 027-046.
Dormann, C. F., Purschke, O., Marquez, J. R. G., Lautenbach, S., & Schroder, B. (2008). Components of uncertainty in species distribution analysis: A case study of the Great Grey Shrike. *Ecology, 89*(12), 3371-3386.
Dormann, C. F., Schymanski, S. J., Cabral, J., Chuine, I., Graham, C., Hartig, F., Kearney, M., Morin, X., Römermann, C., & Schröder, B. (2012). Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography, 39*(12), 2119-2131.
Drescher, J. (2011). The ecology and population structure of the invasive yellow crazy ant *Anoplolepis gracilipes*. *Doktorgrade (Doctor of Philosophy) dissertation. Julius-Maximillians-Universität, Würzburg, Germany,*
Dubois, C. F., & Dubois, A. J. C. (1874). *Les lÉpidoptères de la Belgique [Lepidopteras of Belgium] (J. Feltwell, Trans.)* (Vol. 3): Librairie C. Muquadrt, Merzbach et Falk succrs.
Dukes, J. S., & Mooney, H. A. (1999). Does global change increase the success of biological invaders? *TREE, 14*(4), 135.
Dunning Jr, J. B., Stewart, D. J., Danielson, B. J., Noon, B. R., Root, T. L., Lamberson, R. H., & Stevens, E. E. (1995). Spatially explicit population models: current forms and future uses. *Ecological Applications, 5*(1), 3-11.
Dupin, M., Reynaud, P., Jarošík, V., Baker, R., Brunel, S., Eyre, D., Pergl, J., & Makowski, D. (2011). Effects of the Training Dataset Characteristics on the Performance of Nine Species Distribution Models: Application to *Diabrotica virgifera virgifera*. *PLoS ONE, 6*(6), e20957. doi:10.1371/journal.pone.0020957
Eckert, C. G., Samis, K. E., & Lougheed, S. C. (2008). Genetic variation across species' geographical ranges: the central–marginal hypothesis and beyond. *Molecular Ecology, 17*(5), 1170-1188. doi:10.1111/j.1365-294X.2007.03659.x
Elith, J., Burgman, M. A., & Regan, H. M. (2002). Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecological Modelling, 157*(2-3), 313-329. doi:10.1016/S0304-3800(02)00202-8
Elith, J., & Graham, C. H. (2009). Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography, 32*(1), 66-77.
Elith, J., Graham, C. H., Anderson, R. P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Mortiz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson, A. T., Philips, S. J., Richardson, K., Scachetti-Pereira, R., Schapire, R. E., Soberon, J., Williams, S., Wisz, M. S., & Zimmermann, N. E. (2006). Novel methods improve prediction of species; distributions from occurrence data. *Ecography, 29*(2), 129-151.
Elith, J., Kearney, M., & Phillips, S. (2010). The art of modelling range-shifting species. *Methods in Ecology and Evolution, 1*(4), 330-342. doi:10.1111/j.2041-210X.2010.00036.x
Elith, J., & Leathwick, J. (2007). Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and distributions, 13*(3), 265-275.

Elith, J., & 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., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions, 17*(1), 43-57. doi:10.1111/j.1472-4642.2010.00725.x

Elith, J., Simpson, J., Hirsch, M., & Burgman, M. A. (2013). Taxonomic uncertainty and decision making for biosecurity: spatial models for myrtle/guava rust. *Australasian Plant Pathology, 42*(1), 43-51. doi:10.1007/s13313-012-0178-7

Elton, C. (1958). *The Ecology of Invasions by Animals and Plants*. London: Methuen.

Enders, C. K. (2003). Performing Multivariate Group Comparisons Following a Statistically Significant MANOVA. *Measurement & Evaluation in Counselling & Development (American Counselling Association), 36*(1), 40. Article

Engler, R., Guisan, A., & Rechsteiner, L. (2004). An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology, 41*(2), 263-274.

Epanchin-Niell, R. S., & Hastings, A. (2010). Controlling established invaders: integrating economics and spread dynamics to determine optimal management. *Ecology Letters, 13*(4), 528-541. doi:10.1111/j.1461-0248.2010.01440.x

EPPO. (2011). *Diabrotica virgifera virgifera: Present situation (2011) of Diabrotica virgifera virgifera in Europe. accessed at 03 April 2014*. Retrieved

Er, M., Tunaz, H., & Gökçe, A. (2007). Pathogenicity of entomopathogenic fungi to *Thaumetopoea pityocampa* (Schiff.)(Lepidoptera: Thaumatopoeidae) larvae in laboratory conditions. *Journal of pest science, 80*(4), 235-239.

EROS. (1996). GTOPO30. In E. D. Center (Ed.). Sioux Falls: http://eros.usgs.gov/#/Find_Data/Products_and_Data_Available/gtopo30_info Accessed 2012 Jan 22

Escudero, A., Iriondo, J. M., & Torres, M. E. (2003). Spatial analysis of genetic diversity as a tool for plant conservation. *Biological Conservation, 113*(3), 351-365.

ESRI. (2010). ArcMap (Version 10.0). Redlands, CA: Environmental Systems Research Institute. http://www.esri.com/ Accessed 2011 May 02

Evans, E. A. (2003). Economic Dimensions of Invasive Species. *Choices: the megazine of food, farm and resource issues*(2), 5-10. Retrieved from http://www.choicesmagazine.org/2003-2/2003-2-02.pdf

Ewers, R. M., & Didham, R. K. (2006). Confounding factors in the detection of species responses to habitat fragmentation. *Biological Reviews, 81*(1), 117-142. doi:10.1017/S1464793105006949

Fabre, J. H. (2012). *The Life of the Caterpillar, Volume 6*: Nabu Press. (1916)

Farber, O., & Kadmon, R. (2003). Assessment of alternative approaches for bioclimatic modelling with special emphasis on the Mahalanobis distance. *Ecological Modelling, 160*(1–2), 115-130.

Feltwell, J. (1978). The depredations of the large white butterfly (*Pieris brassicae*)(Pieridae)[Horticultural crops, pests]. *Journal of Research on the Lepidoptera,*

Feltwell, J. (1982). *Large White butterfly: The Biology, Biochemistry, and Physiology of Pieris brassicae (Linnaeus)* (Vol. 18): Springer.

Field, R. P., & Darby, S. M. (1991). Host specificity of the parasitoid, *Sphecophaga vesparum* (Curtis) (Hymenoptera: Ichneumonidae), a potential biological control agent of the social wasps, *Vespula germanica* (Fabricius) and *V. vulgaris* (Linnaeus) (Hymenoptera: Vespidae) in Australia. *New Zealand Journal of Zoology, 18*(2), 193-197. doi:10.1080/03014223.1991.10757966

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation, 24*(01), 38-49. doi:doi:null

Fisher, R. A. (1937). The wave of advance of advantageous genes. *Annals of Eugenics, 7*(4), 355-369. doi:10.1111/j.1469-1809.1937.tb02153.x

fox, J., Friendly, M., & Monette, G. (2013). heplots: Visualizing tests in multivariate linear models (Version R package 1.0-11.). Available from http://CRAN.R-project.org/package=heplots

Frankham, R. (2005). Resolving the genetic paradox in invasive species. *Heredity, 94*(4), 385-385. doi:10.1038/sj.hdy.6800634

Franklin, J. (2010a). *Mapping Species Distributions: Spatial Inference and Prediction*: Cambridge University Press.

Franklin, J. (2010b). Moving beyond static species distribution models in support of conservation biogeography. *Diversity and Distributions, 16*(3), 321-330. doi:10.1111/j.1472-4642.2010.00641.x

Franklin, J., Davis, F. W., Ikegami, M., Syphard, A. D., Flint, L. E., Flint, A. L., & Hannah, L. (2013). Modelling plant species distributions under future climates: how fine scale do climate projections need to be? *Global change biology, 19*(2), 473-483.

Friendly, M., & fox, J. (2013). candisc: Visualizing generalized canonical discriminant and canonical correlation analysis (Version R package 0.6-5). Available from http://CRAN.R-project.org/package=candisc

Frost, T., DeAngelis, D., Bartell, S., Hall, D., & Hurlbert, S. (1988). Scale in the Design and Interpretation of Aquatic Community Research. In S. Carpenter (Ed.), *Complex Interactions in Lake Communities* (pp. 229-258): Springer New York. Retrieved from 10.1007/978-1-4612-3838-6_14. doi:10.1007/978-1-4612-3838-6_14

Furfey, P. H. (1927). A Note on Lefever's "Standard Deviational Ellipse". *American Journal of Sociology, 33*(1), 94-98. doi:10.2307/2765043

Gallien, L., Douzet, R., Pratte, S., Zimmermann, N. E., & Thuiller, W. (2012). Invasive species distribution models – how violating the equilibrium assumption can create new insights. *Global Ecology and Biogeography, 21*(11), 1126-1136. doi:10.1111/j.1466-8238.2012.00768.x

Gardiner, B. O. C. (1995). The Large Cabbage White, *Pieris brassicae*, extends its range to South Africa. *Entomologist's Record and Journal of Variation, 107*(7)

Gardner-Gee, R., & Beggs, J. R. (2013). Invasive wasps, not birds, dominate in a temperate honeydew system. *Austral Ecology, 38*(3), 346-354. doi:10.1111/j.1442-9993.2012.02412.x

Garzon, M. B., Blazek, R., Neteler, M., Dios, R. S. d., Ollero, H. S., & Furlanello, C. (2006). Predicting habitat suitability with machine learning models: The potential area of *Pinus sylvestris* L. in the Iberian Peninsula. *ecological modelling, 197*(3), 383-393.

Gassmann, A. J., Petzold-Maxwell, J. L., Clifton, E. H., Dunbar, M. W., Hoffmann, A. M., Ingber, D. A., & Keweshan, R. S. (2014). Field-evolved resistance by western corn rootworm to multiple *Bacillus thuringiensis* toxins in transgenic maize. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1317179111

Gassmann, A. J., Petzold-Maxwell, J. L., Keweshan, R. S., & Dunbar, M. W. (2011). Field-Evolved Resistance to Bt Maize by Western Corn Rootworm. *PLoS ONE, 6*(7), e22629. doi:10.1371/journal.pone.0022629

Gilbert, M., Fielding, N., Evans, H. F., & Grégoire, J. C. (2003). Spatial pattern of invading *Dendroctonus micans* (Coleoptera: Scolytidae) populations in the United Kingdom. *Canadian Journal of Forest Research, 33*(4), 712-725. doi:10.1139/x02-208

Gilbert, M., Grégoire, J. C., Freise, J. F., & Heitland, W. (2004). Long-distance dispersal and human population density allow the prediction of invasive patterns in the horse chestnut leafminer *Cameraria ohridell*a. *Journal of Animal Ecology, 73*(3), 459-468. doi:10.1111/j.0021-8790.2004.00820.x

Gilbert, M., Guichard, S., Freise, J., GrÉGoire, J. C., Heitland, W., Straw, N., Tilbury, C., & Augustin, S. (2005). Forecasting *Cameraria ohridella* invasion dynamics in recently invaded countries: from validation to prediction. *Journal of Applied Ecology, 42*(5), 805-813. doi:10.1111/j.1365-2664.2005.01074.x

Glare, T. R., Hampton, J. G., Cox, M. P., & Bienkowski, D. A. (2014). United States Patent No. 20140086876 http://www.freepatentsonline.com/y2014/0086876.html.

Gong, J. (2002). Clarifying the standard deviational ellipse. *Geographical Analysis, 34*(2), 155-167.

González, I., & Déjean, S. (2012). CCA: Canonical correlation analysis (Version R package 1.2). Available from http://CRAN.R-project.org/package=CCA

Gorban, A. N. (2007). *Principal Manifolds for Data Visualization and Dimension Reduction* (Vol. 58): Springer.

Gotelli, N. J. (2000). Null model analysis of species co-occurrence patterns. *Ecology, 81*(9), 2606-2621. doi:10.1890/0012-9658(2000)081[2606:NMAOSC]2.0.CO;2

Gottwald, T. R., Hughes, G., Graham, J. H., Sun, X., & Riley, T. (2001). The Citrus Canker Epidemic in Florida: The scientific basis of regulatory eradication policy for an invasive species. *Phytopathology, 91*(1), 31-33.

Gratz, N. G. (2004). Critical review of the vector status of *Aedes albopictus*. *Medical and Veterinary Entomology, 18*(3), 215-227. doi:10.1111/j.0269-283X.2004.00513.x

Gray, M. E., Sappington, T. W., Miller, N. J., Moeser, J., & Bohn, M. O. (2009). Adaptation and invasiveness of Western Corn Rootworm: Intensifying research on a worsening pest. *Annual Review of Entomology, 54*, 305-313. doi:10.1146/annurev.ento.54.110807.090434

Grimm, V. (1999). Ten years of individual-based modelling in ecology: what have we learned and what could we learn in the future? *Ecological Modelling, 115*(2), 129-148. doi:10.1016/S0304-3800(98)00188-4

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., Thulke, H.-H., Weiner, J., Wiegand, T., & DeAngelis, D. L. (2005). Pattern-oriented modelling of agent-based complex systems: lessons from ecology. *science, 310*(5750), 987-991.

Grinnell, J. (1924). Geography and Evolution. *Ecology, 5*(3), 225-229. doi:10.2307/1929447

Gritti, E. S., Gaucherel, C., Crespo-Perez, M.-V., & Chuine, I. (2013). How Can Model Comparison Help Improving Species Distribution Models? *PLoS ONE, 8*(7), e68823. doi:10.1371/journal.pone.0068823

Guisan, A., Edwards Jr, T. C., & Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological modelling, 157*(2), 89-100.

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters, 8*. doi:j.1461-0248.2005.00792.x

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I. T., Regan, T. J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T. G., Rhodes, J. R., Maggini, R., Setterfield, S. A., Elith, J., Schwartz, M. W., Wintle, B. A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M. R., Possingham, H. P., & Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. *Ecology Letters, 16*(12), 1424-1435. doi:10.1111/ele.12189

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling 135*, 150-152. doi:0304-3800:00:$

Gurevitch, J., & Padilla, D. K. (2004). Are invasive species a major cause of extinctions? *Trends in ecology & evolution (Personal edition), 19*(9), 470-474.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research, 3*, 1157-1182.

Hagstrum, D. W., & Subramanyam, B. (2010). Immature Insects: Ecological Roles of Mobility. *American Entomologist, 56*(4), 231.

Haines, I. H., & Haines, J. B. (1979). Toxic bait for the control of *Anoplolepis longipes* (Jerdon) (Hymenoptera: Formicidae) in the Seychelles. II. Effectiveness, specificity and cost of baiting in field applications. *Bulletin of Entomological Research, 69*(01), 77-85. 10.1017/S0007485300017909

Hanberry, B. B., He, H. S., & Palik, B. J. (2012). Pseudoabsence Generation Strategies for Species Distribution Models. *PLoS ONE, 7*(8), e44486. doi:10.1371/journal.pone.0044486

Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M., & Dougherty, E. R. (2010). Small-sample precision of ROC-related estimates. *Bioinformatics, 26*(6), 822-830. doi:10.1093/bioinformatics/btq037

Hannah, L., Midgley, G. F., & Millar, D. (2002). Climate change-integrated conservation strategies. *Global Ecology and Biogeography, 11*(6), 485-495. doi:10.1046/j.1466-822X.2002.00306.x

Hansen, T., & Merivee, E. (1971). Cold-hardiness of the cabbage butterflies *Pieris brassicae* L. and *Pieris rapae* L. *Akad Nauk Est Ssr Izv Ser Biol, 1971*

Hanski, I., & Gilpin, M. (1991). Metapopulation dynamics: Brief history and conceptual domain. *Biological Journal of the Linnean Society, 42*, 5-10.

Hanson, S. M., & Craig Jr, G. B. (1995). Relationship between cold hardiness and supercooling point in *Aedes albopictus* eggs. *Journal of the American Mosquito Control Association, 11*(1), 35-38.

Harris, R. J. (1991). Diet of the wasps *Vespula vulgaris* and *V. germanica* in honeydew beech forest of the South Island, New Zealand. *New Zealand Journal of Zoology, 18*(2), 159-169. doi:10.1080/03014223.1991.10757963

Harris, R. J., Harcourt, S. J., Glare, T. R., Rose, E. A. F., & Nelson, T. J. (2000). Susceptibility of *Vespula vulgaris* (Hymenoptera: Vespidae) to Generalist Entomopathogenic Fungi and Their Potential for Wasp Control. *Journal of Invertebrate Pathology, 75*(4), 251-258. doi:10.1006/jipa.2000.4928
Hartley, S., Harris, R., & Lester, P. J. (2006). Quantifying uncertainty in the potential distribution of an invasive species: climate and the Argentine ant. *Ecology Letters, 9*(9), 1068-1079. doi:10.1111/j.1461-0248.2006.00954.x
Hastie, T., & Fithian, W. (2013). Inference from presence-only data; the ongoing controversy. *Ecography*, 864–867. doi:10.1111/j.1600-0587.2013.00321.x
Hastie, T., & Stuetzle, W. (1989). Principal Curves. *Journal of the American Statistical Association, 84*(406), 502-516. doi:10.1080/01621459.1989.10478797
Hastings, A., Cuddington, K., Davies, K. F., Dugaw, C. J., Elmendorf, S., Freestone, A., Harrison, S., Holland, M., Lambrinos, J., Malvadkar, U., Melbourne, B. A., Moore, K., Taylor, C., & Thomson, D. (2005). The spatial spread of invasions: new developments in theory and evidence. *Ecology Letters, 8*(1), 91-101. doi:10.1111/j.1461-0248.2004.00687.x

Hawley, W. A. (1988). The biology of *Aedes albopictus*. *Journal of the American Mosquito Control Association. Supplement, 1*, 1.
Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation* (2 ed.): Prentice Hall.
Heath, J. (1970). Provisional atlas of the insects of the British Isles. Part 1: Lepidoptera Rhopalocera. Butterflies (Maps 1 to 57).

Heikkinen, R. K., Luoto, M., Kuussaari, M., & Pöyry, J. (2005). New insights into butterfly–environment relationships using partitioning methods. *Proceedings of the Royal Society B: Biological Sciences, 272*(1577), 2203-2210. doi:10.1098/rspb.2005.3212
Held, C., & Spieth, H. R. (1999). First evidence of pupal summer diapause in *Pieris brassicae* L.: the evolution of local adaptedness. *Journal of Insect Physiology, 45*(6), 587-598. doi:10.1016/S0022-1910(99)00042-6
Helmuth, B., Kingsolver, J. G., & Carrington, E. (2005). Biophysics, physiological ecology, and climate change: Does Mechanism Matter? *Annual Review of Physiology, 67*(1), 177-201. doi:doi:10.1146/annurev.physiol.67.040403.105027
Hengl, T. (2009). *A Practical Guide to Geostatistical Mapping*. Open Access Publication: Available: http://spatial-analyst.net/book/system/files/Hengl_2009_GEOSTATe2c1w.pdf Accessed 2012 April 15.
Henmerik, L., Busstra, C., & Mols, P. (2004). Predicting the temprature-dependent natural population expansion of the western corn rootworm, *Diabrotica virgifera*. *Entomologia Expermimentalis et Applicata, 111*. doi:j.0013-8703.2004.00150.x

Higgins, S. I., & Richardson, D. M. (1999). Predicting plant migration rates in a changing world: the role of long-distance dispersal. *The American Naturalist, 153*(5), 464-475.
Higgins, S. I., Richardson, D. M., & Cowling, R. M. (1996). Modeling Invasive Plant Spread: The Role of Plant-Environment Interactions and Model Structure. *Ecology, 77*(7), 2043-2054. doi:10.2307/2265699
Hijmans, R. J., Cameron, S. E., & Parra, J. L. (2005a). *BIOCLIM*. Retrieved from http://www.worldclim.org/bioclim
Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005b). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology, 25*(15), 1965-1978. doi:10.1002/joc.1276
Hill, M., Holm, K., Vel, T., Shah, N. J., & Matyot, P. (2003). Impact of the introduced yellow crazy ant *Anoplolepis gracilipes* on Bird Island, Seychelles. *Biodiversity & Conservation, 12*(9), 1969-1984.
Hirzel, A. H., Hausser, J., Chessel, D., & Perrin, N. (2002). Ecological-Niche factor analysis: How to compute habitat-suitability maps without absence data? *Ecology, 83*(7), 2027.
Hirzel, A. H., Heifer, V., & Metral, F. (2001). Assessing habitat-suitability models with a virtual species. *Ecological Modelling, 145*, 111-121.
Hirzel, A. H., & Le Lay, G. (2008). Habitat suitability modelling and niche theory. *Journal of Applied Ecology, 45*(5), 1372-1381. doi:10.1111/j.1365-2664.2008.01524.x
Hochberg, M. E. (1991). Intra-host interactions between a braconid endoparasitoid, *Apanteles glomeratus*, and a baculovirus for larvae of *Pieris brassicae*. *The Journal of Animal Ecology*, 51-63.
Hódar, J. A., Castro, J., & Zamora, R. (2003). Pine processionary caterpillar *Thaumetopoea pityocampa* as a new threat for relict Mediterranean Scots pine forests under climatic warming. *Biological Conservation, 110*(1), 123-129.
Hoffmann, B. D. (2014). Quantification of supercolonial traits in the yellow crazy ant, *Anoplolepis gracilipes*. *Journal of Insect Science (Madison), 14*(25)
Hoffmann, B. D., & O'Connor, S. (2004). Eradication of two exotic ants from Kakadu National Park. *Ecological Management & Restoration, 5*(2), 98-105.
Holzmann, H., & Vollmer, S. (2008). A likelihood ratio test for bimodality in two-component mixtures with application to regional income distribution in the EU. *AStA Advances in Statistical Analysis, 92*(1), 57-69. doi:10.1007/s10182-008-0057-2

Honório, N. A., Silva, W. d. C., Leite, P. J., Gonçalves, J. M., Lounibos, L. P., & Lourenço-de-Oliveira, R. (2003). Dispersal of *Aedes aegypti* and *Aedes albopictus* (Diptera: Culicidae) in an urban endemic dengue area in the State of Rio de Janeiro, Brazil. *Memórias do Instituto Oswaldo Cruz, 98*, 191-198.
Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *biometrical journal, 50*(3), 346-363.
Hothorn, T., Hornik, K., VanDerWiel, M. A., & Zeilleis, A. (2006a). A lego system for conditional inference. *The American Statistician, 60*(3), 257-263.
Hothorn, T., Hornik, K., & Zeileis, A. (2006b). Unbiased recursive partitioning: A conditional inference framework. *Journal of Comput. Graph. Stat., 15*(3), 651-674.
Howell, D. (2007). Statistical methods for psychology Thomson Wadsworth. *Belmont, CA*, 1-739.
Hulme, P. E. (2006). Beyond control: wider implications for the management of biological invasions. *Journal of Applied Ecology, 43*, 835-845.
Hulme, P. E. (2009). Trade, transport and trouble: managing invasive species pathways in an era of globalization. *Journal of Applied Ecology, 46*. doi:10.1111/j.1365-2664.2008.01600.x

Hummel, H. E., Dinnesen, S., Nedelev, T., Modic, S., Urek, G., & Ulrichs, C. (2008). *Dibrotica virgifera virgifera* LeConte in confrontation mood: simultaneous geographical and host spectrum expansion in southeastern Slovenia. *Mitt.Dtsch. Ges.Allg. Ent, 16*, 127-128.
Hutchinson, G. E. (1957). Concluding Remarks. *Cold Spring Harbor Symposia on Quantitative Biology, 22*, 415-427. doi:10.1101/sqb.1957.022.01.039
Hutchinson, G. E. (1978). An introduction to population biology. New Hawen: Yale University Press
Jackson, J. J., & Brooks, M. A. (1995). Parasitism of western corn rootworm larvae and pupae by *Steinernema carpocapsae*. *Journal of nematology, 27*(1), 15.
Jaffe, G. (2003). Planting trouble: Are farmers squandering *Bt* corn technology. *Center for Science in the Public Interest, Washington, DC, available at: http://cspinet.org/new/pdf/bt_corn_report.pdf,*
Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR), 31*(3), 264-323.

Jiménez-Valverde, A. (2012). Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography, 21*(4), 498-507. doi:10.1111/j.1466-8238.2011.00683.x
Jiménez-Valverde, A., & Lobo, J. M. (2006). The ghost of unbalanced species distribution data in geographical model predictions. *Diversity and Distributions, 12*(5), 521-524. doi:10.1111/j.1366-9516.2006.00267.x
Jiménez-Valverde, A., & Lobo, J. M. (2007). Threshold criteria for conversion of probability of species presence to either–or presence–absence. *Acta Oecologica, 31*(3), 361-369. doi:10.1016/j.actao.2007.02.001
Jiménez-Valverde, A., Lobo, J. M., & Hortal, J. (2008). Not as good as they seem: the importance of concepts in species distribution modelling. *Diversity and Distributions, 14*(6), 885-890. doi:10.1111/j.1472-4642.2008.00496.x
Johnson, M. L., & Gaines, M. S. (1990). Evolution of dispersal: theoretical models and empirical tests using birds and mammals. *Annual review of ecology and systematics*, 449-480.
Jolivet, P. (1992). *Insects and Plants: Parallel Evolution & Adaptations, Second Edition* (2nd ed.): Taylor & Francis.
Jorgensen, S. E. (1986). *Fundamentals of Ecological Modelling* (Vol. 9). Copenhagen, Denmark: Elsevier.
Kampichler, C., Wieland, R., Calmé, S., Weissenberger, H., & Arriaga-Weiss, S. (2010). Classification in conservation biology: A comparison of five machine-learning methods. *Ecological Informatics, 5*(6), 441-450. doi:10.1016/j.ecoinf.2010.06.003
Kanat, M., Alma, M. H., & Sivrikaya, F. (2005). Effect of defoliation by *Thaumetopoea pityocampa* (Den. & Schiff.)(Lepidoptera: Thaumetopoeidae) on annual diameter increment of *Pinus brutia* Ten. in Turkey. *Annals of forest science, 62*(1), 91-94.
Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). Kernlab-An S4 package for kernel methods in R. *Journal of statistical software, 11*(9), 1-20.
Kean, J. (2013). How accurate are thermal accumulation methods for predicting phenology in New Zealand. *New Zealand Plant Protection, 66*, 124-131.
Kean, J., & Phillips, C. (2013a). *OR or AND and risk mapping*. Lincoln, New Zealand.
Kean, J., & Phillips, C. (2013b, 2 July 2013). *Phenology and diapause research for great white butterfly (Pieris brassicae)* (No. RE400/2012/445). Lincoln, New Zealand: agresearch. (Report for Ministry for Primary Industries)
Kean, J., & Phillips, C. (2013c, 2 July 2013). *Phenology and diapause research for great white butterfly (Pieris brassicae)* (No. RE400/2012/447). Lincoln, New Zealand: agresearch. (Report for Ministry for Primary Industries)
Kean, J., & Phillips, C. (2013d, 10 October 2013). *Phenology and diapause research for great white butterfly (Pieris brassicae)* (No. RE400/ 2013/481). Lincoln, New Zealand: agresearch. (Report for Ministry for Primary Industries)
Kearney, M. (2006). Habitat, environment and niche: what are we modelling? *Oikos, 115*(1), 186-191. doi:10.1111/j.2006.0030-1299.14908.x
Kearney, M., Phillips, B. L., Tracy, C. R., Christian, K. A., Betts, G., & Porter, W. P. (2008). Modelling species distributions without using species distributions: the cane toad in Australia under current and future climates. *Ecography, 31*(4), 423-434. doi:10.1111/j.0906-7590.2008.05457.x
Kearney, M., & Porter, W. (2009). Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters, 12*, 334. doi:10.1111/j.1461-0248.2008.01277.x
Kearney, M., Wintle, B. A., & Porter, W. P. (2010). Correlative and mechanistic models of species distribution provide congruent forecasts under climate change. *Conservation Letters, 3*(3), 203-213. doi:10.1111/j.1755-263X.2010.00097.x
Keeling, M. J., Woolhouse, M. E. J., Shaw, D. J., Matthews, L., Chase-Topping, M., Haydon, D. T., Cornell, S. J., Kappey, J., Wilesmith, J., & Grenfell, B. T. (2001). Dynamics of the 2001 UK Foot and Mouth Epidemic: Stochastic Dispersal in a Heterogeneous Landscape. *Science, 294*(5543), 813-817. doi:10.1126/science.1065973
Keith, J. M., & Spring, D. (2013). Agent-based Bayesian approach to monitoring the progress of invasive species eradication programs. *Proceedings of the National Academy of Sciences, 110*(33), 13428-13433. doi:10.1073/pnas.1216146110

Kellner, C. V., & Shapiro, A. M. (1983). Ecological Interactions of *Pieris brassicae* L. (Lepidoptera: Pieridae) and Native Pierini in Chile. *Studies on Neotropical Fauna and Environment, 18*(1), 53-64. doi:10.1080/01650528309360618
Kenis, M., Auger-Rozenberg, M.-A., Roques, A., Timms, L., Péré, C., Cock, M. W., Settele, J., Augustin, S., & Lopez-Vaamonde, C. (2009). Ecological effects of invasive alien insects. *Biological Invasions, 11*(1), 21-45. doi:10.1007/s10530-008-9318-y
Kerr, J. T., & Ostrovsky, M. (2003). From space to species: Ecological applications for remote sensing. *TRENDS in ecology and Evolution, 18*(6), 299, 300. doi:10.1016/S0169-5347(03)00071-5
Kjær, C., Damgaard, C., & Lauritzen, A. (2009). Assessment of effects of Bt-oilseed rape on large white butterfly (*Pieris brassicae*) in natural habitats.
Klein, H. Z. (1932). Studien zur Ökologie und Epidemiologie der Kohlweißlinge. *Zeitschrift für Angewandte Entomologie, 19*(3), 395-448. doi:10.1111/j.1439-0418.1932.tb00316.x
Kleinbaum, D. G., & Klein, M. (2005). *Logistic Regression* (2 ed.). New York: Springer.
Koehler, A., & Geisenhoffer, C. (2012). *NAPPFAST Model Documentation; Pieris brassicae developmental requirements*.

Kokko, H., & Lopez-Sepulcre, A. (2010). From individual dispersal to species ranges: Perspectives for a changing world *Science 313*, 789-790. doi:10.1126/science.1128566
Kolar, C. S., & Lodge, D. M. (2001). Progress in invasion biology: predicting invaders. *Trends in Ecology & Evolution, 16*(4), 199-204. doi:10.1016/S0169-5347(01)02101-2
Kot, M., Lewis, M. A., & Driessche, P. v. d. (1996). Dispersal Data and the Spread of Invading Organisms. *Ecology, 77*(7), 2027-2042. doi:10.2307/2265698
Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal, 37*(2), 233-243.
Kriticos, D. J., Le Maitre, D. C., & Webber, B. L. (2013). Essential elements of discourse for advancing the modelling of species' current and potential distributions. *Journal of Biogeography, 40*(3), 608-611. doi:10.1111/j.1365-2699.2012.02791.x

Kriticos, D. J., Reynaud, P., Baker, R., & Eyre, D. (2012a). Estimating the global area of potential establishment for the western corn rootworm (*Diabrotica virgifera virgifera*) under rain-fed and irrigated agriculture*. *EPPO Bulletin, 42*(1), 56-64.

Kriticos, D. J., Webber, B. L., Leriche, A., Ota, N., Macadam, I., Bathols, J., & Scott, J. K. (2012b). CliMond: global high-resolution historical and future scenario climate surfaces for bioclimatic modelling. *Methods in Ecology and Evolution, 3*(1), 53-64. doi:10.1111/j.2041-210X.2011.00134.x
Labbé, G. M. C., Scaife, S., Morgan, S. A., Curtis, Z. H., & Alphey, L. (2012). Female-Specific Flightless (fsRIDL) Phenotype for Control of *Aedes albopictus*. *PLoS Negl Trop Dis, 6*(7), e1724. doi:10.1371/journal.pntd.0001724
Lambrechts, L., Scott, T. W., & Gubler, D. J. (2010). Consequences of the Expanding Global Distribution of *Aedes albopictus* for Dengue Virus Transmission. *PLoS Negl Trop Dis, 4*(5), e646. doi:10.1371/journal.pntd.0000646
Larsen, T. B. (1974). *Butterflies of Lebanon*: National Council for Scientific Research.
Lawler, J. J., White, D., Neilson, R. P., & Blaustein, A. R. (2006). Predicting climate-induced range shifts: model differences and model reliability. *Global Change Biology, 12*(8), 1568-1584.
Lee, C. E. (2002). Evolutionary genetics of invasive species. *Trends in Ecology & Evolution 17*(8), 386-391.
Lefever, D. W. (1926). Measuring Geographic Concentration by Means of the Standard Deviational Ellipse. *American Journal of Sociology, 32*(1), 88-94. doi:10.2307/2765249
Lefort, M. C., Barratt, B. I. P., Marris, J. W. M., & Boyer, S. (2012). Combining molecular and morphological approaches to differentiate the pest *Costelytra zealandica* (White) (Coleoptera: Scarabeidae: Melolonthinae) from the non-pest *Costelytra brunneum* (Broun) at the larval stage. *New Zealand Entomologist, 36*(1), 15-21. doi:10.1080/00779962.2012.742369
Legendre, P., & Legendre, L. (2012). *Numerical Ecology* (Vol. 20): Elsevier.
Lester, P. J., & Tavite, A. (2004). Long-legged ants, *Anoplolepis gracilipes* (Hymenoptera: Formicidae), have invaded Tokelau, changing composition and dynamics of ant and invertebrate communities. *Pacific Science, 58*(3)
Levin, S. A. (1992). The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. *Ecology, 73*(6), 1943-1967. doi:10.2307/1941447
Levine, E., & Oloumi-Sadeghi, H. (1996). Western Corn Rootworm (Coleoptera: Chrysomelidae) Larval Injury to corn grown for seed production following soybeans grown for seed production. *J. Econ. Entomol., 89*(4). doi:0022-0493/96/1010-1016$02.00/0
Li, W., Guo, Q., & Elkan, C. (2011). Can we model the probability of presence of species without absence data? *Ecography, 34*(6), 1096-1105. doi:10.1111/j.1600-0587.2011.06888.x
Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest: http://cran.r-project.org/web/packages/randomForest/index.html Accessed 2012 Jun 13
Lippitt, C. D., Rogan, J., Toledano, J., Sangermano, F., Eastman, J. R., Mastro, V., & Sawyer, A. (2008). Incorporating anthropogenic variables into a species distribution model to map gypsy moth risk. *Ecological Modelling, 210*, 340-343. doi:10.1016/j.ecolmodel.2007.08.005
Lizée, M.-H., Mauffrey, J.-F., Tatoni, T., & Deschamps-Cottin, M. (2011). Monitoring urban environments on the basis of biological traits. *Ecological Indicators, 11*(2), 353-361. doi:10.1016/j.ecolind.2010.06.003
Lobo, J. M. (2008). More complex distribution models or more representative data? *Biodiversity Informatics, 5*, 14-19. istribution models, model reliability, pseudo-absences, conservation usefulness.
Lobo, J. M., Jiménez-Valverde, A., & Hortal, J. (2010). The uncertain nature of absences and their importance in species distribution modelling. *Ecography, 33*(1), 103-114. doi:10.1111/j.1600-0587.2009.06039.x
Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography, 17*(2), 145-151. doi:10.1111/j.1466-8238.2007.00358.x
Lobo, J. M., Verdú, J. R., & Numa, C. (2006). Environmental and geographical factors affecting the Iberian distribution of flightless *Jekelius* species (Coleoptera: Geotrupidae). *Diversity and Distributions, 12*(2), 179-188.
Lopez-Osorio, F., Pickett, K. M., Carpenter, J. M., Ballif, B. A., & Agnarsson, I. (2014). Phylogenetic relationships of yellowjackets inferred from nine loci (Hymenoptera: *Vespidae, Vespinae, Vespula* and *Dolichovespula*). *Molecular Phylogenetics and Evolution, 73*(0), 190-201. doi:10.1016/j.ympev.2014.01.007
Lorena, A. C., Jacintho, L. F. O., Siqueira, M. F., Giovanni, R. D., Lohmann, L. G., de Carvalho, A. C. P. L. F., & Yamamoto, M. (2011). Comparing machine learning classifiers in potential distribution modelling. *Expert Systems with Applications, 38*(5), 5268-5275.
Lounibos, L. P. (2002). Invasions by insect vectors of human disease *Annual Review of Entomology, 47*(1), 233-266. doi:doi:10.1146/annurev.ento.47.091201.145206
Luoto, M., Virkkala, R., Heikkinen, R. K., & Rainio, K. (2004). Predicting bird species richness using remote sensing in boreal agricultural-forest mosaics. *Ecological Applications, 14*(6), 1946-1962. doi:10.1890/02-5176
Lütolf, M., Kienast, F., & Guisan, A. (2006). The ghost of past species occurrence: improving species distribution models for presence-only data. *Journal of Applied Ecology, 43*(4), 802-815. doi:10.1111/j.1365-2664.2006.01191.x
Maboudou-Tchao, E. M., & Agboto, V. (2013). Monitoring the covariance matrix with fewer observations than variables. *Computational Statistics & Data Analysis, 64*(0), 99-112. doi:10.1016/j.csda.2013.02.028

MacNally, R. (2000). Regression and model-building in conservation biology, biogeography and ecology: The distinction between – and reconciliation of – 'predictive' and 'explanatory' models. *Biodiversity & Conservation, 9*(5), 655-671. doi:10.1023/A:1008985925162
Mader, H. J., Schell, C., & Kornacker, P. (1990). Linear barriers to Arthropod movements in the landscape. *Biological conservation, 54*, 219-221.
MAF. (2012). Great white cabbage butterfly. In M. o. A. a. Forestry (Ed.) (pp. 2): MAF http://www.biosecurity.govt.nz/files/pests/great-white-cabbage-butterfly-fact-sheet.pdf
Manel, S., Dias, J. M., Buckton, S. T., & Ormerod, S. J. (1999). Alternative methods for predicting species distribution: an illustration with Himalayan river birds. *Journal of Applied Ecology, 36*(5), 734-747. doi:10.1046/j.1365-2664.1999.00440.x
Manel, S., Williams, H. C., & Ormerod, S. J. (2001). Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology, 38*(5), 921-931. doi:10.1046/j.1365-2664.2001.00647.x
Manevitz, L. M., & Yousef, M. (2002). One-class svms for document classification. *J. Mach. Learn. Res., 2*, 139-154. doi:944808
Marivate, V. N., Nelwamodo, F. V., & Marwala, T. (2007). Autoencoder, principal component analysis and support vector regression for data imputation. *arXiv preprint arXiv:0709.2506,*
Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R. K., & Thuiller, W. (2009). Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions, 15*(1), 59-69. doi:10.1111/j.1472-4642.2008.00491.x
Masciocchi, M., Beggs, J. R., Carpenter, J. M., & Corley, J. C. (2010). Primer registro de *Vespula vulgaris* (Hymenoptera: Vespidae) en la Argentina. *Revista de la Sociedad Entomológica Argentina, 69*(3-4), 267-270.
MathWorks. (2011). MATLAB (Version 7.12.0.635). Massachusetts: The MathWorks Inc. http://www.mathworks.com.au Accessed 2012 Jun 13
Matthews, R. W., Goodisman, M. A., Austin, A. D., & Bashford, R. (2000). The introduced English wasp *Vespula vulgaris* (L.)(Hymenoptera: Vespidae) newly recorded invading native forests in Tasmania. *Australian Journal of Entomology, 39*(3), 177-179.
McCallum, A., & Nigam, K. (1998). *A comparison of event models for naive bayes text classification.* Paper presented at the AAAI-98 workshop on learning for text categorization
McGeoch, M. A., Butchart, S. H. M., Spear, D., Marais, E., Kleynhans, E. J., Symes, A., chanson, J., & Hoffmann, M. (2010). Global indicators of biological invasion: species numbers, biodiversity impact and policy responses. *Diversity and Distributions, 16*, 95. doi:10.1111/j.1472-4642.2009.00633.x
McPherson, J. M., Jetz, W., & Rogers, D. J. (2004). The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology, 41*(5), 811-823. doi:10.1111/j.0021-8901.2004.00943.x
Medley, K. A. (2009). Niche shifts during the global invasion of the Asian tiger mosquito, *Aedes albopictus* Skuse (Culicidae), revealed by reciprocal distribution models. *Global Ecology and Biogeography, 19*(1), 122-133.
Medlock, J. M., Avenell, D., Barrass, I., & Leach, S. (2006). Analysis of the potential for survival and seasonal activity of *Aedes albopictus* (Diptera: Culicidae) in the United Kingdom. *Journal of Vector Ecology, 31*(2), 292-304.
Mendiburu, F. d. (2012). agricolae: statistical procedures for agricultural research R package version 1.1-2. http://CRAN.R-project.org/package=agricolae Accessed 2012 Sep 12. Retrieved Accessed 2012 Sep 12
Meyer, D., Dimitriadou, E., honik, K., Leisch, F., & Weingessel, A. (2007). e1071: Misc functions of the department of statistics (e1071) (Version R package 1.5-17): http://cran.r-project.org/web/packages/e1071/index.html Accessed 2012 Jun 13. Retrieved Accessed 2012 Sep 12
MFE. (2004). LCDB2. In M. f. t. Environment (Ed.), *LCDB* https://koordinates.com/login/?next=/layer/1072-land-cover-database-version-2-lcdb2/
Miller, N., Estoup, A., Toepfer, S., Bourguet, D., Lapchin, L., Derridj, S., Kim, K. S., reynaud, P., Furlan, L., & Guillemaud, T. (2005). Muliple transatlantic introductions of the western corn rootworm. *Science, 310*, 992. doi:10.1126/science.1115871
Miller, N. A., & Stillman, J. H. (2012). Physiological optima and critical limits. *Nature Education Knowledge, 3*(10), 1.
Mitchell, A. (2005). *Spatial Measurements and Statistics* (Vol. 2): ESRI Press.
Moeser, J., & Vidal, S. (2004). Do alternative host plants enhance the invasion of maize pest *Diabrotica virgifera virgifera* (coleoptera: Chrysomelidae, Galerucinae) in Europe. *Environmental Entomology, 33*(5), 1170, 1174-1176. doi:0046-225X/04/1169Ð1177$04.00/0
Monahan, W. B. (2009). A Mechanistic Niche Model for Measuring Species' Distributional Responses to Seasonal Temperature Gradients. *PLoS ONE, 4*(11), e7921. doi:10.1371/journal.pone.0007921
Monahan, W. B., & Tingley, M. W. (2012). Niche Tracking and Rapid Establishment of Distributional Equilibrium in the House Sparrow Show Potential Responsiveness of Species to Climate Change. *PLoS ONE, 7*(7), e42097. doi:10.1371/journal.pone.0042097
Mooney, H. A., & Cleland, E. E. (2001). The evolutionary impact of invasive species. *Proceedings of the National Academy of Sciences, 98*(10), 5446-5451. doi:10.1073/pnas.091093398
Morency, L.-P., Kok, I., & Gratch, J. (2010). A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems, 20*(1), 70-84. doi:10.1007/s10458-009-9092-y
Morin, X., & Lechowicz, M. J. (2008). Contemporary perspectives on the niche that can improve models of species range shifts under climate change. *Biology Letters, 4*(5), 573-576. doi:10.1098/rsbl.2008.0181
Morisette, J. T., Jarnevich, C. S., Ullah, A., Cai, W., Pedelty, J. A., Gentle, J. E., Stohlgren, T. J., & Schnase, J. L. (2006). A tamarisk habitat suitability map for the continental United States. *Frontiers in Ecology and the Environment, 4*(1), 11-17. doi:10.1890/1540-9295
Morozov, A., Ruan, S., & Li, B.-L. (2008). Patterns of patchy spread in multi-species reaction–diffusion models. *Ecological Complexity, 5*(4), 313-328. doi:10.1016/j.ecocom.2008.05.002
Nabout, J. C., Caetano, J. M., Ferreira, R. B., Teixeira, I. R., & Alves, S. M. d. F. (2012). Using correlative, mechanistic and hybrid niche models to predict the productivity and impact of global climate change on maize crop in Brazil. *Brazillian journal of nature conservation, 10*(2), 177. doi:10.4322/natcon.2012.034

NAPPO. (2002). Pathway identified for introduction of *Pieris brassicae* (Lepidoptera: Pieridae) from Europe. Massachusssetts: NAPPO http://www.pestalert.org/viewArchNewsStory.cfm?nid=205&keyword=pieris%20brassicae
NASA-GSFC. (2000). *EOS data products handbook* PDF. Greenbelt, Maryland: National Aeronautics and Space Administration (NASA) Goddard Space Flight Center. http://eospso.gsfc.nasa.gov/ftp_docs/data_products_vol2.pdf Accessed 2010 Dec 22.
Nathan, R., & Muller-Landau, H. C. (2000). Spatial patterns of seed dispersal, their determinants and consequences for recruitment. *Trends in Ecology & Evolution, 15*(7), 278-285. doi:10.1016/S0169-5347(00)01874-7
Neteler, M., Bowman, M. H., Landa, M., & Metz, M. (2012). GRASS GIS: A multi-purpose open source GIS. *Environmental Modelling & Software, 31*(0), 124-130. doi:10.1016/j.envsoft.2011.11.014
Neubert, M. G., Kot, M., & Lewis, M. A. (1995). Dispersal and Pattern Formation in a Discrete-Time Predator-Prey Model. *Theoretical Population Biology, 48*(1), 7-43. doi:10.1006/tpbi.1995.1020
Neubert, M. G., & Parker, I. M. (2004). Projecting Rates of Spread for Invasive Species. *Risk Analysis, 24*(4), 817-831. doi:10.1111/j.0272-4332.2004.00481.x
Nishimatsu, T., & Jackson, J. J. (1998). Interaction of Insecticides, Entomopathogenic Nematodes, and Larvae of the Western Corn Root worm (Coleoptera: Chrysomelidae). *Journal of Economic Entomology, 91*(2), 410-418.
NIWA. (2005). 30 years daily minimum and maximum temperature for New Zealand: NIWA
Nix, H. A. (1986). A biogeographic analysis of Australian elapid snakes. . In R. Longmore (Ed.), *Australian Flora and Fauna Series* (Atlas of Elapid Snakes of Australia ed.). Canberra: Australian Government Publishing Service
Novacek, M. J., & Cleland, E. E. (2001). *The current biodiversity extinction event: Scenarios for mitigation and recovery*. Paper presented at the The National Academy of Sciences Colloquium.
O'Dowd, D. J., Green, P. T., & Lake, P. (1999). *Status, Impact, and Recommendations for Research and Management of Exotic Invasive Ants in Christmas Island National Park*. Clayton, Victoria: Centre for Analysis and Management of Biological Invasions, Monash University.
O'Dowd, D. J., Green, P. T., & Lake, P. S. (2003). Invasional 'meltdown'on an oceanic island. *Ecology Letters, 6*(9), 812-817.

Odum, E. P. (1971). *Fundamentals of Ecology*: Saunders.
Økland, B., Skarpaas, O., Schroeder, M., Magnusson, C., Lindelöw, Å., & Thunes, K. (2010). Is Eradication of the Pinewood Nematode (*Bursaphelenchus xylophilus*) Likely? An Evaluation of Current Contingency Plans. *Risk Analysis, 30*(9), 1424-1439. doi:10.1111/j.1539-6924.2010.01431.x
Okubo, A., & Simon, A. L. (1989). A Theoretical Framework for Data Analysis of Wind Dispersal of Seeds and Pollen. *Ecology, 70*(2), 329-338. doi:10.2307/1937537
Onstad, D. W., Crwoder, D. W., Isard, S. A., Levine, E., Spencer, J. L., O'neal, M. E., Ratcliffe, S. T., Gray, M. E., Bledsoe, L. W., Fonzo, C. D. D., Eisley, J. B., & Edwards, C. R. (2003). Does landscape diversity slow the spread of rotation-resistant Western Corn Rootworm (Coleoptera: Chyrysomelidae)? *Environmental Entomology, 32*(5). doi:0046-225X/03/0992Ð1001$04.00/0
Onstad, D. W., Guse, C. A., Spencer, J. L., Levine, E., & Gray, M. E. (2001). Modelling the dynamics of adaptation to transgenic corn by western corn rootworm (Coleoptera: chrysomelidae). *Jounal of Econ. Entomology, 94*(2). doi:0022-0493/01/0529Ð0540$02.00/0
Onstad, D. W., Joselyn, M. G., Isard, S. A., Levine, E., Spencer, J. L., Bledsoe, L. W., Edwards, C. R., Fonzo, C. D. D., & Willson, H. (1999). Modelling the spread of Western Corn Rootworm (Coleoptera: Chrysomelidae) populations adapting to Soyabean-Corn rotation. *Environmental Entomology, 28*(2), 191-193.
Paini, D. R., Worner, S. P., Cook, D. C., De Barro, P. J., & Thomas, M. B. (2010). Threat of invasive pests from within national borders. *Nat Commun*, 115. doi:10.1038/ncomms1118
Pascoe, A. (2002). Crazy ant species found. *Stowaways, 12*(2)
Pasek, J. E. (1988). Influence of wind and windbreaks on local dispersal of insects. *Agriculture, Ecosystems & Environment, 22–23*(0), 539-554. doi:10.1016/0167-8809(88)90044-8
Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2*(11), 559-572.
Pearson, R. G., Dawson, T. P., & Liu, C. (2004). Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. *Ecography, 27*(3), 285-298. doi:10.1111/j.0906-7590.2004.03740.x
Pearson, R. G., Thuiller, W., Araújo, M. B., Martinez-Meyer, E., Brotons, L., McClean, C., Miles, L., Segurado, P., Dawson, T. P., & Lees, D. C. (2006). Model-based uncertainty in species range prediction. *Journal of Biogeography, 33*(10), 1704-1711. doi:10.1111/j.1365-2699.2006.01460.x
Pebesma, E. J., & Bivand, R. S. (2005). SP: Classes and methods for spatial data in R: Accessed 2012 Jun 13 http://cran.r-project.org/web/packages/sp/index.html
Pejchar, L., & Mooney, H. A. (2009). Invasive species, ecosystem services and human well-being. *Trends in ecology & evolution (Personal edition), 24*(9), 497-504.
Pereira, J. M. C., & Itami, R. M. (1991). GIS-based habitat modelling using logistic multiple regression: a study of the Mt. Graham red squirrel. *Photogrammetric Engineering & Remote Sensing, 57*(11), 1476,1482. doi:0099-1112/91/5711-1475$03.00/0
Peterson, A. T. (2006). Uses and Requirements of Ecological Niche Models and Related Distributional Models. *Biodiversity Informatics, 3*, 59-72.
Phillips, C., Brown, K., Green, C., Walker, G., Broome, K., Toft, R., Lee, B. V., & King, M. (2013). *Great white butterfly interim report prepared for Ministry for Primary Industries external technical advisory group*.
Phillips, C., Vink, C. J., Blanchet, A., & Hoelmer, K. A. (2008). Hosts are more important than destinations: What genetic variation in *Microctonus aethiopoides* (Hymenoptera: Braconidae) means for foreign exploration for natural enemies. *Molecular phylogenetics and evolution, 49*(2), 467-476.
Phillips, S. J. (2008). Transferability, sample selection bias and background data in presence only modelling: a response to paterson et al. (2007). *Ecography, 31*, 272. doi:10.1111/j.2007.0906-7590.05378.x
Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling, 190*(3–4), 231-259.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications, 19*(1), 181-197. doi:10.1890/07-2153.1

Pimentel, C., Calvao, T., Santos, M., Ferreira, C., Neves, M., & Nilsson, J.-Å. (2006). Establishment and expansion of a *Thaumetopoea pityocampa* (Den. & Schiff.)(Lep. Notodontidae) population with a shifted life cycle in a production pine forest, Central-Coastal Portugal. *Forest ecology and management, 233*(1), 108-115.

Pimentel, D., Zuniga, R., & Morrison, D. (2004). Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecological Economics, 52*, 273, 282. doi:10.1016/j.ecolecon.2004.10.002

Pimm, S. L., & Rice, J. C. (1987). The dynamics of multispecies, multi-life-stage models of aquatic food webs. *Theoretical Population Biology, 32*(3), 303-325. doi:10.1016/0040-5809(87)90052-9

Pitt, J. P. W. (2008). *Modelling the Spread of Invasive Species Across Heterogeneous Landscapes.* Lincoln University, Lincoln.

Pitt, J. P. W., Kriticos, D. J., & Dodd, M. B. (2011). Temporal limits to simulating the future spread pattern of invasive species: *Buddleja davidii* in Europe and New Zealand. *Ecological Modelling, 222*(11), 1880-1887. doi:10.1016/j.ecolmodel.2011.03.023

Pitt, J. P. W., Worner, S. P., & Suarez, A. V. (2009). Predicting Argentine ant spread over the heterogeneous landscape using a spatially explicit stochastic model. *Ecological Applications, 19*(5), 1176-1186. doi:10.1890/08-1777.1

Plećaš, M., Gagić, V., Janković, M., Petrović-Obradović, O., Kavallieratos, N. G., Tomanović, Ž., Thies, C., Tscharntke, T., & Ćetković, A. (2014). Landscape composition and configuration influence cereal aphid–parasitoid–hyperparasitoid interactions and biological control differentially across years. *Agriculture, Ecosystems & Environment, 183*(0), 1-10. doi:10.1016/j.agee.2013.10.016

Poulos, H. M., Chernoff, B., Fuller, P. L., & Butman, D. (2012). Ensemble forecasting of potential habitat for three invasive fishes. *Aquatic Invasions, 7*(1), 59-72. doi:10.3391/ai.2012.7.01

Pulliam, H. R., Dunning, J. B., & Liu, J. (1992). Population Dynamics in Complex Landscapes: A Case Study. *Ecological Applications, 2*(2), 165-177. doi:10.2307/1941773

Pullin, A. S., & Bale, J. S. (1989). Influence of diapause and temperature on cryoprotectant synthesis and cold hardiness in pupae of *Pieris brassicae*. *Comparative Biochemistry and Physiology Part A: Physiology, 94*(3), 499-503.

Pullin, A. S., Bale, J. S., & Fontaine, X. L. R. (1991). Physiological aspects of diapause and cold tolerance during overwintering in *Pieris brassicae*. *Physiological Entomology, 16*(4), 447-456. doi:10.1111/j.1365-3032.1991.tb00584.x

R Core Team. (2012). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/ Accessed 2012 Oct 29

Rangi, D. (2009). Invasive species and poverty: the missing link... *Environment matters,* 12-13.

Raymond, B., McInnes, J., Dambacher, J. M., Way, S., & Bergstrom, D. M. (2011). Qualitative modelling of invasive species eradication on Subantarctic Macquarie Island. *Journal of Applied Ecology, 48*(1), 181-191. doi:10.1111/j.1365-2664.2010.01916.x

Reschenhofer, E. (2001). The bimodality principle. *J Stat Educ, 9*(1), 1-16.

Rich, P. M., & Weiss, S. B. (1991). *Spatial models of microclimate and habitat suitability: Lessons from threatened species.* Paper presented at the the eleventh annual ESRI user conference, Palm Springs, CA

Richards-Zawacki, C. L. (2009). Effects of slope and riparian habitat connectivity on gene flow in an endangered panamanian frog, *Atelopus varius*. *Diversity and Distributions, 15*, 796. doi:10.1111/j.1472-4642.2009.00582.x

Richardson, D. M., & Pyšek, P. (2008). Fifty years of invasion ecology – the legacy of Charles Elton. *Diversity and Distributions, 14*(2), 161-168. doi:10.1111/j.1472-4642.2007.00464.x

Richardson, D. M., Pyšek, P., Rejmánek, M., Barbour, M. G., Panetta, F. D., & West, C. J. (2000). Naturalization and invasion of alien plants: concepts and definitions. *Diversity and distributions, 6*(2), 93-107.

Ripley, B. D. (1994). Neural Networks and Related Methods for Classification. *Journal of the Royal Statistical Society. Series B (Methodological), 56*(3), 409-456. doi:10.2307/2346118

Robinet, C., Baier, P., Pennerstorfer, J., Schopf, A., & Roques, A. (2007). Modelling the effects of climate change on the potential feeding activity of *Thaumetopoea pityocampa* (Den. & Schiff.) (Lep., Notodontidae) in France. *Global Ecology and Biogeography, 16*(4), 460-471. doi:10.1111/j.1466-8238.2006.00302.x

Robinet, C., Imbert, C.-E., Rousselet, J., Sauvard, D., Garcia, J., Goussard, F., & Roques, A. (2012). Human-mediated long-distance jumps of the pine processionary moth in Europe. *Biological Invasions, 14*(8), 1557-1569.

Robinet, C., Roques, A., Pan, H., Fang, G., Ye, J., Zhang, Y., & Sun, J. (2009). Role of Human-mediated dispersal in the spread of the Pinewood Nematode in China. *PLoS One, 4*(2), 1. doi:10.1371/journal.pone.0004646

Rodda, G. H., Jarnevich, C. S., & Reed, R. N. (2011). Challenges in Identifying Sites Climatically Matched to the Native Ranges of Animal Invaders. *PLoS ONE, 6*(2), e14670. doi:10.1371/journal.pone.0014670

Rödder, D., Schmidtlein, S., Veith, M., & Lötters, S. (2009). Alien Invasive Slider Turtle in Unpredicted Habitat: A Matter of Niche Shift or of Predictors Studied? *PLoS ONE, 4*(11), e7843. doi:10.1371/journal.pone.0007843

Roiz, D., Neteler, M., Castellani, C., Arnoldi, D., & Rizzoli, A. (2011). Climatic Factors Driving Invasion of the Tiger Mosquito (*Aedes albopictus*) into New Areas of Trentino, Northern Italy. *PLoS ONE, 6*(4), e14800.

Rothschild, M., & Schoonhoven, L. (1977). Assessment of egg load by *Pieris brassicae* (Lepidoptera: Pieridae).

Roura-Pascual, N., Brotons, L., Peterson, A., & Thuiller, W. (2009). Consensual predictions of potential distributional areas for invasive species: a case study of Argentine ants in the Iberian Peninsula. *Biological Invasions, 11*(4), 1017-1031. doi:10.1007/s10530-008-9313-3

Roura-Pascual, N., Suarez, A. V., McNyset, K., Gómez, C., Pons, P., Touyama, Y., Wild, A. L., Gascon, F., & Peterson, A. T. (2006). Niche differentiation and fine-scale projections for Argentine ants based on remotely sensed data. *Ecological Applications, 16*(5), 1832-1841.

Rousselet, J., Zhao, R., Argal, D., Simonato, M., Battisti, A., Roques, A., & Kerdelhué, C. (2010). The role of topography in structuring the demographic history of the pine processionary moth, *Thaumetopoea pityocampa* (Lepidoptera: Notodontidae). *Journal of biogeography, 37*(8), 1478-1490.

Rowland, E., Davison, J., & Graumlich, L. (2011). Approaches to Evaluating Climate Change Impacts on Species: A Guide to Initiating the Adaptation Planning Process. *Environmental Management, 47*(3), 322-337. doi:10.1007/s00267-010-9608-x

Ruckelshaus, M., Hartway, C., & Kareiva, P. (1997). Assessing the Data Requirements of Spatially Explicit Dispersal Models. *Conservation Biology, 11*(6), 1298-1306. doi:10.1046/j.1523-1739.1997.96151.x

Ruscoe, W. A., Ramsey, D. S. L., Pech, R. P., Sweetapple, P. J., Yockney, I., Barron, M. C., Perry, M., Nugent, G., Carran, R., Warne, R., Brausch, C., & Duncan, R. P. (2011). Unexpected consequences of control: competitive vs. predator release in a four-species assemblage of invasive mammals. *Ecology Letters, 14*(10), 1035-1042. doi:10.1111/j.1461-0248.2011.01673.x

Salvato, P., Battisti, A., Concato, S., Masutti, L., Patarnello, T., & Zane, L. (2002). Genetic differentiation in the winter pine processionary moth (*Thaumetopoea pityocampa—wilkinsoni* complex), inferred by AFLP and mitochondrial DNA markers. *Molecular Ecology, 11*(11), 2435-2444.

Sangermano, F., & Eastman, R. (2007). *Linking GIS and ecology–the use of Mahalanobis typicalities to model species distribution.* Paper presented at the Memorias XI Conferencia Iberoamericana de Sistemas de Información Geográfica

Saunders, J. T., Saunders, C. M., Buwalda, J. G., Gerard, P. J., Bourdôt, G. W., Wratten, S. D., & Goldson, S. L. (2013). The economic impact of failures in plant protection to New Zealand. *PeerJ PrePrints, 1*, e140v141. doi:10.7287/peerj.preprints.140v1

Saura, S. (2002). Effects of minimum mapping unit on land cover data spatial configuration and composition. *International Journal of Remote Sensing, 23*(22), 4853-4880.

Saura, S., & Martinez-Millan, J. (2001). Sensitivity of landscape pattern metrics to map spatial extent. *Photogrammetric engineering and remote sensing, 67*(9), 1027-1036.

Savage, D., & Renton, M. (2014). Requirements, design and implementation of a general model of biological invasion. *Ecological Modelling, 272*(0), 394-409. doi:10.1016/j.ecolmodel.2013.10.001

Sax, D. F., Stachowicz, J. J., Brown, J. H., Bruno, J. F., Dawson, M. N., Gaines, S. D., Grosberg, R. K., Hastings, A., Holt, R. D., Mayfield, M. M., O'Connor, M. I., & Rice, W. R. (2007). Ecological and evolutionary insights from species invasions. *Trends in ecology & evolution, 22*(9), 465-471. doi:10.1016/j.tree.2007.06.009

Schaepman, M. E., Ustin, S. L., Plaza, A. J., Painter, T. H., Verrelst, J., & Liang, S. (2009). Earth system science related imaging spectroscopy—An assessment. *Remote Sensing of Environment, 113, Supplement 1*(0), S123-S137. doi:10.1016/j.rse.2009.03.001

Schellhorn, N. A., Bianchi, F. J. J. A., & Hsu, C. L. (2014). Movement of Entomophagous Arthropods in Agricultural Landscapes: Links to Pest Suppression. *Annual Review of Entomology, 59*(1), 559-581. doi:10.1146/annurev-ento-011613-161952

Schoenberg, I. (1971). On equidistant cubic spline interpolation. *Bulletin of the American Mathematical Society, 77*(6), 1039-1044.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation, 13*(7), 1443-1471.

Schölkopf, B., Smola, A., & Müller, K.-R. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation, 10*(5), 1299-1319. doi:10.1162/089976698300017467

Scholte, E. J., Dijkstra, E., Blok, H., De Vries, A., Takken, W., Hofhuis, A., Koopmans, M., De Boer, A., & Reusken, C. B. E. M. (2008). Accidental importation of the mosquito *Aedes albopictus* into the Netherlands: a survey of mosquito distribution and the presence of dengue virus. *Medical and Veterinary Entomology, 22*(4), 352-358. doi:10.1111/j.1365-2915.2008.00763.x

Scholte, E. J., & Schaffner, F. (2007). 14. Waiting for the tiger: establishment and spread of the *Aedes albopictus* mosquito in Europe. *Emerging pests and vector-borne diseases in Europe, 1*, 241.

Scholz, M., Fraunholz, M., & Selbig, J. (2008). Nonlinear Principal Component Analysis: Neural Network Models and Applications. In A. N. Gorban (Ed.), *Principal manifolds for data visualization and dimension reduction* (pp. 44-67): Springer.

Scholz, M., & Vigario, R. (2002). Nonlinear PCA: a new hierarchical approach.

Sebastien, A., Gruber, M. A. M., & Lester, P. J. (2012). Prevalence and genetic diversity of three bacterial endosymbionts (*Wolbachia, Arsenophonus*, and *Rhizobiales*) associated with the invasive yellow crazy ant (*Anoplolepis gracilipes*). *Insectes Sociaux, 59*(1), 33-40. doi:10.1007/s00040-011-0184-8

Segurado, P., & Araújo, M. B. (2004). An evaluation of methods for modelling species distributions. *Journal of Biogeography, 31*(10), 1555-1568. doi:10.1111/j.1365-2699.2004.01076.x

Senay, S. D., Worner, S. P., & Ikeda, T. (2013). Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modelling. *PLoS ONE, 8*(8), e71218. doi:10.1371/journal.pone.0071218

Sharov, A. A., Roberts, E. A., & Liebhold, A. M. (1995). Gypsy Moth (lepidoptera: Lymantriidae) Spread in the central Appalachians: Three methods for species boundary estimation. *Environmental Entomology, 24*(6), 1530.

Shigesada, N., Kawasaki, K., & Takeda, Y. (1995). Modelling Stratified Diffusion in Biological Invasions. *The American Naturalist, 146*(2), 229-251. doi:10.2307/2463059

Sillero, N. (2011). What does ecological modelling model? A proposed classification of ecological niche models based on their underlying methods. *Ecological Modelling, 222*(8), 1343-1346. doi:10.1016/j.ecolmodel.2011.01.018

Simberloff, D. (1996). Impacts of introduced species in the United States. *Consequences, 2*

Simberloff, D. (2006). Invasional meltdown 6 years later: important phenomenon, unfortunate metaphor, or both? *Ecology Letters, 9*(8), 912-919.

Simberloff, D. (2011). Charles Elton: Neither Founder nor Siren, but Prophet. In D. M. Richardson (Ed.), *Fifty years of invasion ecology: The legacy of Charles Elton* (1st ed., pp. 11-24): Wiley-blackwell.

Sinclair, S. J., White, M. D., & Newell, G. R. (2010). How useful are species distribution models for managing biodiversity under future climates? . *Ecology and Society, 15*(1), 8.

Skellam, J. G. (1951). Random dispersal in theoretical populations. *Biometrika, 38*(1-2), 196-218. doi:10.1093/biomet/38.1-2.196

Skelsey, P., With, K. A., & Garrett, K. A. (2013). Pest and Disease Management: Why We Shouldn't Go against the Grain. *PLoS ONE, 8*(9), e75892. doi:10.1371/journal.pone.0075892

Skuse, F. A. A. (1895). The banded mosquito of Bengal. *Indian Museum Notes, 3 (for 1894)*(5)

Soberón, J. (2007). Grinnellian and Eltonian niches and geographic distributions of species. *Ecology letters, 10*(12), 1115-1123.

Soberón, J., & Peterson, A. T. (2005). *Interpretation of Models of Fundamental Ecological Niches and Species' Distributional Areas* (Vol. 2).

Sømme, L. (1967). The effect of temperature and anoxia on haemolymph composition and supercooling in three overwintering insects. *Journal of Insect Physiology, 13*(5), 805-814. doi:10.1016/0022-1910(67)90128-X

Spencer, J. L., Isard, S. A., & Levine, E. (1999). Free flight of Western Corn Rootworm (Coleoptrera: Chrysomelidae) to corn and soybean plants in a walk-in wind tunnel. *J. Econ. Entomol., 92*(1), 153-154. doi:0022-0493/99/0146Ð0155$02.00/0

Spieth, H. R. (2002). Estivation and hibernation of *Pieris brassicae* (L.) in southern Spain: synchronization of two complex behavioural patterns. *Population Ecology, 44*(3), 0273-0280. doi:10.1007/s101440200031

Spieth, H. R., & Cordes, R. (2012). Geographic comparison of seasonal migration events of the large white butterfly, *Pieris brassicae*. *Ecological Entomology, 37*(6), 439-445. doi:10.1111/j.1365-2311.2012.01385.x

Spieth, H. R., & Kaschuba-Holtgrave, A. (1996). A new experimental approach to investigate migration in *Pieris brassicae* L. *Ecological Entomology, 21*(3), 289-294. doi:10.1111/j.1365-2311.1996.tb01246.x

Spieth, H. R., Pörschmann, U., & Teiwes, C. (2011). The occurrence of summer diapause in the large white butterfly *Pieris brassicae* (Lepidoptera: Pieridae): A geographical perspective. *European Journal of Entomology, 108*, 377-384.

Spieth, H. R., & Sauer, K. P. (1991). Quantitative measurement of photoperiods and its significance for the induction of diapause in *Pieris brassicae* (Lepidoptera, Pieridae). *Journal of Insect Physiology, 37*(3), 231-238. doi:10.1016/0022-1910(91)90073-9

Spradbery, J. P. (1973). *Wasps. An Account of The Biology and Natural History of Social and Solitary Wasps, with Particular Reference to Those of The British Isles*.

Spurr, E. B. (1995). Protein bait preferences of wasps (*Vespula vulgaris* and *V. germanica*) at Mt Thomas, Canterbury, New Zealand. *New Zealand Journal of Zoology, 22*(3), 281-289. doi:10.1080/03014223.1995.9518043

Stastny, M., Battisti, A., Petrucco-Toffolo, E., Schlyter, F., & Larsson, S. (2006). Host-plant use in the range expansion of the pine processionary moth, *Thaumetopoea pityocampa*. *Ecological entomology, 2006*

Steckel, J., Westphal, C., Peters, M. K., Bellach, M., Rothenwoehrer, C., Erasmi, S., Scherber, C., Tscharntke, T., & Steffan-Dewenter, I. (2014). Landscape composition and configuration differently affect trap-nesting bees, wasps and their antagonists. *Biological Conservation, 172*(0), 56-64. doi:10.1016/j.biocon.2014.02.015

Steiner, A. (2010). counting the cost of alien invasions. *BBC*. Retrieved from http://news.bbc.co.uk/2/hi/science/nature/8615398.stm

Steyerberg, E. W., Eijkemans, M. J. C., & Habbema, J. D. F. (1999). Stepwise Selection in Small Data Sets: A Simulation Study of Bias in Logistic Regression Analysis. *Journal of Clinical Epidemiology, 52*(10), 935-942. doi:10.1016/S0895-4356(99)00103-1

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

Stockwell, D., & Peterson, A. T. (2002). Effects of sample size on accuracy of species distribution models. *Ecological Modelling, 148*(1), 1-13. doi:10.1016/S0304-3800(01)00388-X

Suckling, D., Barrington, A., Chhagan, A., Stephens, A., Burnip, G., Charles, J., & Wee, S. (2007). Eradication of the Australian Painted Apple Moth *Teia anartoides* in New Zealand: Trapping, Inherited Sterility, and Male Competitiveness. In *Area-wide control of insect pests* (pp. 603-615): Springer.

Sutherst, R. W. (2000). Climate Change and Invasive Species: a Conceptual Framework. In *Invasive species in a changing world* (pp. 211-240).

Sutherst, R. W. (2014). Pest species distribution modelling: origins and lessons from history. *Biological Invasions, 16*(2), 239-256.

Sutherst, R. W., & Maywald, G. F. (1985). A computerised system for matching climates in ecology. *Agriculture, Ecosystems & Environment, 13*(3–4), 281-299. doi:10.1016/0167-8809(85)90016-7

Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics* (4th ed ed.). New York: HarperCollins.

Tabashnik, B. E., Brevault, T., & Carriere, Y. (2013). Insect resistance to Bt crops: lessons from the first billion acres. *Nat Biotech, 31*(6), 510-521. Research. doi:10.1038/nbt.2597

Thomas, C., Moller, H., Plunkett, G., & Harris, R. (1990). The prevalence of introduced *Vespula vulgaris* wasps in a New Zealand beech forest community. *New Zealand journal of ecology, 13*(1), 63-72.

Thuiller, W. (2003). BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology, 9*(10), 1353-1362. doi:10.1046/j.1365-2486.2003.00666.x

Thuiller, W. (2004). Patterns and uncertainties of species' range shifts under climate change. *Global Change Biology, 10*(12), 2020-2027. doi:10.1111/j.1365-2486.2004.00859.x

Thuiller, W., Brotons, L., Araújo, M. B., & Lavorel, S. (2004). Effects of restricting environmental range of data to project current and future species distributions. *Ecography, 27*(2), 165-172. doi:10.1111/j.0906-7590.2004.03673.x

Tingley, M. W., Monahan, W. B., Beissinger, S. R., & Moritz, C. (2009). Birds track their Grinnellian niche through a century of climate change. *Proceedings of the National Academy of Sciences, 106*(Supplement 2), 19637-19643. doi:10.1073/pnas.0901562106

Toepfer, S., Gueldenzoph, C., Ehlers, R. U., & Kuhlmann, U. (2005). Screening of entomopathogenic nematodes for virulence against the invasive western corn rootworm, *Diabrotica virgifera virgifera* (Coleoptera: Chrysomelidae) in Europe. *Bulletin of Entomological Research, 95*(05), 473-482. 10.1079/BER2005379

Toepfer, S., & Kuhlmann, U. (2005). Natural Mortality Factors Acting on Western Corn Rootworm Populations: a Comparison Between the United States and Central Europe. In S. Vidal, U. Kuhlmann & C. R. Edwards (Eds.), *Western corn rootworm [electronic resource]: ecology and management*: CABI.

Toepfer, S., Levay, N., & Kiss, J. (2006). Adult movements of newly introduced alien *Diabrotica virgifera virgifera* (Coleoptera: Chrysomelidae) from non-host habitats. *Bulletin of Entomological Research, 96*. doi:10.1079/BER2006430

Toft, R. J., & Harris, R. J. (2004). Can trapping control Asian paper wasp(Polistes chinensis antennalis) populations? *New Zealand Journal of Ecology, 28*(2), 279-282.

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D., & Kadmon, R. (2007). A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions, 13*(4), 397-405. doi:10.1111/j.1472-4642.2007.00346.x

Tsutsui, N. D., Suarez, A. V., Holway, D. A., & Case, T. J. (2000). Reduced genetic variation and the success of an invasive species. *Proceedings of the National Academy of Sciences, 97*(11), 5948-5953. doi:10.1073/pnas.100110397

Turner, M. G., Gardner, R. H., & O'Neill, R. V. (2001). *Landscape Ecology in Theory and Practice: Pattern and Process*: U.S. Government Printing Office.

Turnock, W. J., & Fields, P. G. (2005). Winter climates and coldhardiness in terrestrial insects. *European Journal of Entomology, 102*(4), 561-576.

Tuv, E., Borisov, A., Runger, G., & Torkkola, K. (2009). Feature selection with ensembles, artificial variables, and redundancy elimination. *The Journal of Machine Learning Research, 10*, 1341-1366.

UchmaDski, J., & Grimm, V. (1996). Individual-based modelling in ecology: what makes the difference? *Trends in ecology & evolution (Personal edition), 11*(10), 437-441.

Urban, M. C., Phillips, B. L., Skelly, D. K., & Shine, R. (2007). The cane toad's (*Chaunus [Bufo] marinus*) increasing ability to invade Australia is revealed by a dynamically updated range model. *Proc Biol Sci, 274*(1616), 1413-1419. doi:10.1098/rspb.2007.0114

Urban, M. C., Phillips, B. L., Skelly, D. K., & Shine, R. (2008). A Toad More Travelled: The Heterogeneous Invasion Dynamics of Cane Toads in Australia. *The American Naturalist, 171*(3), E134-E148. doi:10.1086/527494

Valle, M., van Katwijk, M. M., de Jong, D. J., Bouma, T. J., Schipper, A. M., Chust, G., Benito, B. M., Garmendia, J. M., & Borja, Á. (2013). Comparing the performance of species distribution models of *Zostera marina*: Implications for conservation. *Journal of Sea Research, 83*, 56-64.

VanDerWal, J., Shoo, Luke P., Graham, C., & Williams, Stephen E. (2009). Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecological Modelling, 220*(24), 589-594. doi:10.1016/j.ecolmodel.2008.11.010

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Venables, W. N., & Ripley, B. D. (1997). *Modern Applied Statistics With S-plus* (2 ed.). New York: Springer-Verlag.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics With S.* (4 ed.). New York: springer.

Venette, R. C., Kriticos, D. J., Magarey, R. D., Koch, F. H., Baker, R. H. A., Worner, S. P., Raboteaux, N. N. G., McKenney, D. W., Dobesberger, E. J., Yemshanov, D., Barro, P. J. D., Hutchison, W. D., Fowler, G., Kalaris, T. M., & Pedlar, J. (2010). Pest Risk Maps for Invasive Alien Species: A Roadmap for Improvement. *BioScience, 60*(5), 349-362.

Vink, C. J., Derraik, J. G. B., Phillips, C. B., & Sirvid, P. J. (2010). The invasive Australian redback spider, *Latrodectus hasseltii* Thorell 1870 (Araneae: Theridiidae): current and potential distributions, and likely impacts. *Biological Invasions*. doi:10.1007/s10530-010-9885-6

Waage, J. K., & Mumford, J. D. (2008). Agricultural biosecurity. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1492), 863-876. doi:10.1098/rstb.2007.2188

Walker, P. A., & Cocks, K. D. (1991). HABITAT: A Procedure for Modelling a Disjoint Environmental Envelope for a Plant or Animal Species. *Global Ecology and Biogeography Letters, 1*(4), 108-118.

Walsh, C., & Nally, R. M. (2013). hier.part: Hierarchical Partitioning (Version R package 1.0-4). Available from http://CRAN.R-project.org/package=hier.part

Wang, G., Gertner, G. Z., Fang, S., & Anderson, A. (2005). A methodology for spatial uncertainty analysis of remote sensing and GIS products. *Photogrammetric engineering and remote sensing, 71*(12), 1423.

Warton, D. I., & Shepherd, L. C. (2010). Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *Ann. Appl. Stat., 4*(3), 1383-1402. doi:10.1214/10-AOAS331

Watts, M. J., & Worner, S. P. (2008). Comparing ensemble and cascaded neural networks that combine biotic and abiotic variables to predict insect species distribution. *Ecological Informatics, 3*(6), 354-356. doi:10.1016/j.ecoinf.2008.08.003

Way, M. J., Scargle, J. D., Ali, K. M., & Srivastava, A. N. (2012). *Advances in Machine Learning and Data Mining for Astronomy*: Taylor & Francis.

Weihe, P. E., & Neely, R. K. (1997). The effects of shading on competition between purple loosestrife and broad-leaved cattail. *Aquatic botany, 59*, 130-131.

Weihs, C., Ligges, U., Luebke, K., & Raabe, N. (2005). klaR Analyzing German Business Cycles. In D. Baier, R. Decker & L. Schmidt-thieme (Eds.), *Data analysis and decision support* (pp. 335-343). Berlin: Springer-Verlag.

Wetterer, J. K. (2005). Worldwide Distribution and Potential Spread of the Long-Legged Ant, *Anoplolepis gracilipes* (Hymenoptera: Formicidae). *Sociobiology, 45*(1), 77-97.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Wigglesworth, V. B. (1945). Transpiration Through the Cuticle of Insects. *Journal of Experimental Biology, 21*(3-4), 97-114.

Williams, C. B. (1958). *Insect Migration* (Vol. 36). London: Collins.

Wisz, M., & Guisan, A. (2009). Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. *BMC Ecology, 9*(1), 8.

Worldbank. (2007). *Invasive species can cause serious ecological, environmental and health impact, says World Bank*. Retrieved from http://climatechange.worldbank.org/node/3844

Worner, S. P. (1988). Evaluation of Diurnal Temperature Models and Thermal Summation in New Zealand. *Journal of Economic Entomology, 81*(1), 9-13.

Worner, S. P. (1991). Use of Models in Applied Entomology: The Need for Perspective. *Environmental Entomology, 20*(3), 768-773.

Worner, S. P. (1992). Performance of Phenological Models Under Variable Temperature Regimes: Consequences of the Kaufmann or Rate Summation Effect. *Environmental Entomology, 21*(4), 689-699.

Worner, S. P. (1994). Predicting The Establishment of Exotic Pests in Relation to Climate. In J. L. Sharp & G. J. Hallman (Eds.), *Quarantine treatments for pests of food plants* (pp. 1132). Boulder, Colorado, USA: Westview Press.

Worner, S. P., & Gevrey, M. (2006). Modelling global insect pest species assemblages to determine risk of invasion. *Journal of Applied Ecology, 43*(5), 858-867. doi:10.1111/j.1365-2664.2006.01202.x

Worner, S. P., Gevrey, M., Ikeda, T., Leday, G., Pitt, J., Schliebs, S., & Soltic, S. (2014). Ecological Informatics for the Prediction and Management of Invasive Species. In N. Kasabov (Ed.), *Springer Handbook of Bio-/Neuroinformatics* (pp. 565-583): Springer Berlin Heidelberg. Retrieved from 10.1007/978-3-642-30574-0_35. doi:10.1007/978-3-642-30574-0_35

Worner, S. P., Ikeda, T., Leday, G., & Joy, M. (2010). *Surveillance tools for freshwater invertebrates*. (2010/21)

Wyss, J. H. (2000). Screwworm Eradication in the Americas. *Annals of the New York Academy of Sciences, 916*(1), 186-193. doi:10.1111/j.1749-6632.2000.tb05289.x

Yan, M., Cao, W., Luo, W., & Jiang, H. (2000). A mechanistic model of phasic and phenological development of wheat. I. Assumption and description of the model. *The journal of applied ecology, 11*(3), 355.

Yemshanov, D., Koch, F. H., Barry Lyons, D., Ducey, M., & Koehler, K. (2012). A dominance-based approach to map risks of ecological invasions in the presence of severe uncertainty. *Diversity and Distributions, 18*(1), 33-46. doi:10.1111/j.1472-4642.2011.00848.x

Yesson, C., Taylor, M. L., Tittensor, D. P., Davies, A. J., Guinotte, J., Baco, A., Black, J., Hall-Spencer, J. M., & Rogers, A. D. (2012). Global habitat suitability of cold-water octocorals. *Journal of Biogeography, 39*(7), 1278-1292. doi:10.1111/j.1365-2699.2011.02681.x

Zaniewski, A. E., Lehmann, A., & Overton, J. M. (2002). Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling, 157*(2–3), 261-280.

Zapata, F., & Jiménez, I. (2012). Species Delimitation: Inferring Gaps in Morphology across Geography. *Systematic Biology, 61*(2), 179-194. doi:10.1093/sysbio/syr084

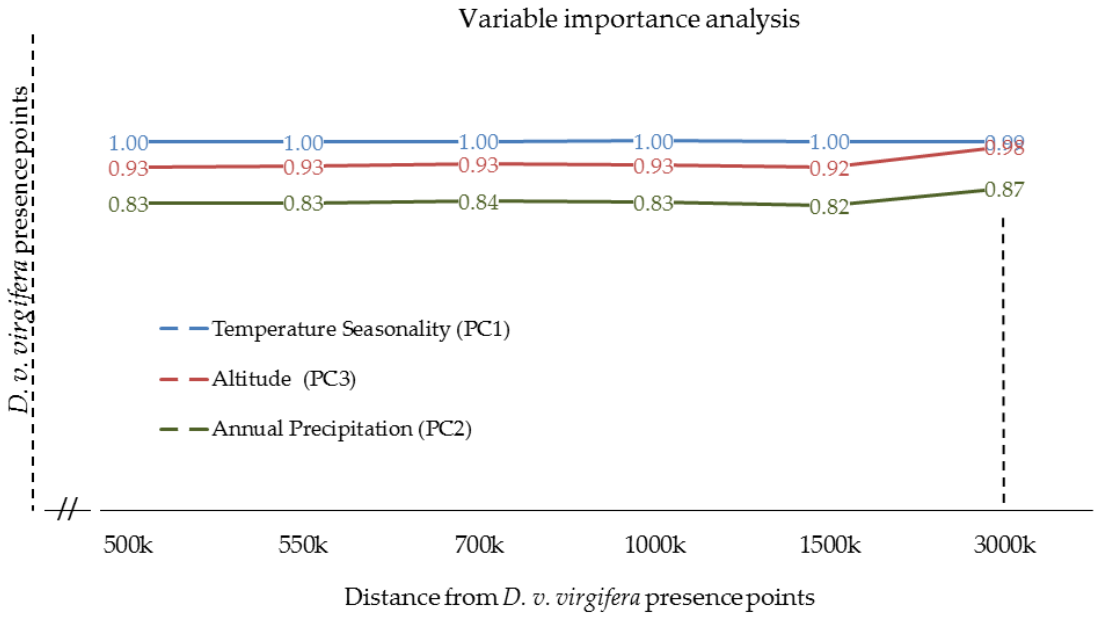
Zavaleta, E. S., Hobbs, R. J., & Mooney, H. A. (2001). Viewing invasive species removal in a whole-ecosystem context. *Trends in Ecology & Evolution, 16*(8), 454-459.

Zimmermann, N. E., JR, T. C. E., Moisen, G. G., Frescino, T. S., & Blackard, J. A. (2007). Remote sensing-based predictors improve distribution models of rare, early successional and broadleaf tree species in Utah. *Journal of Applied Ecology, 44*, 1058-1060. doi:10.1111/j.1365-2664.2007.01348.x

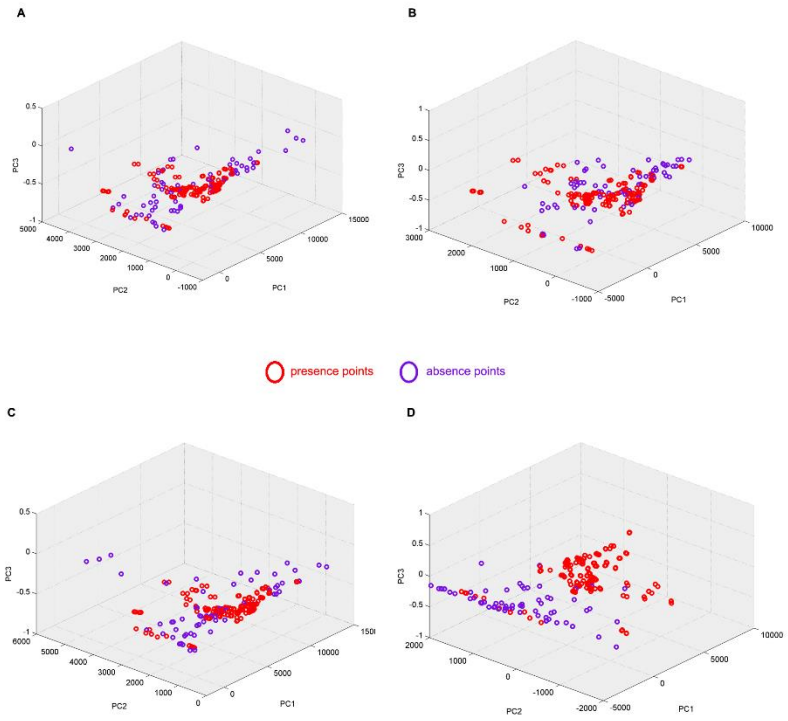
Appendices

10. Appendices

Appendix 3.1 Pseudo-absence generation for *D. v. virgifera* case study

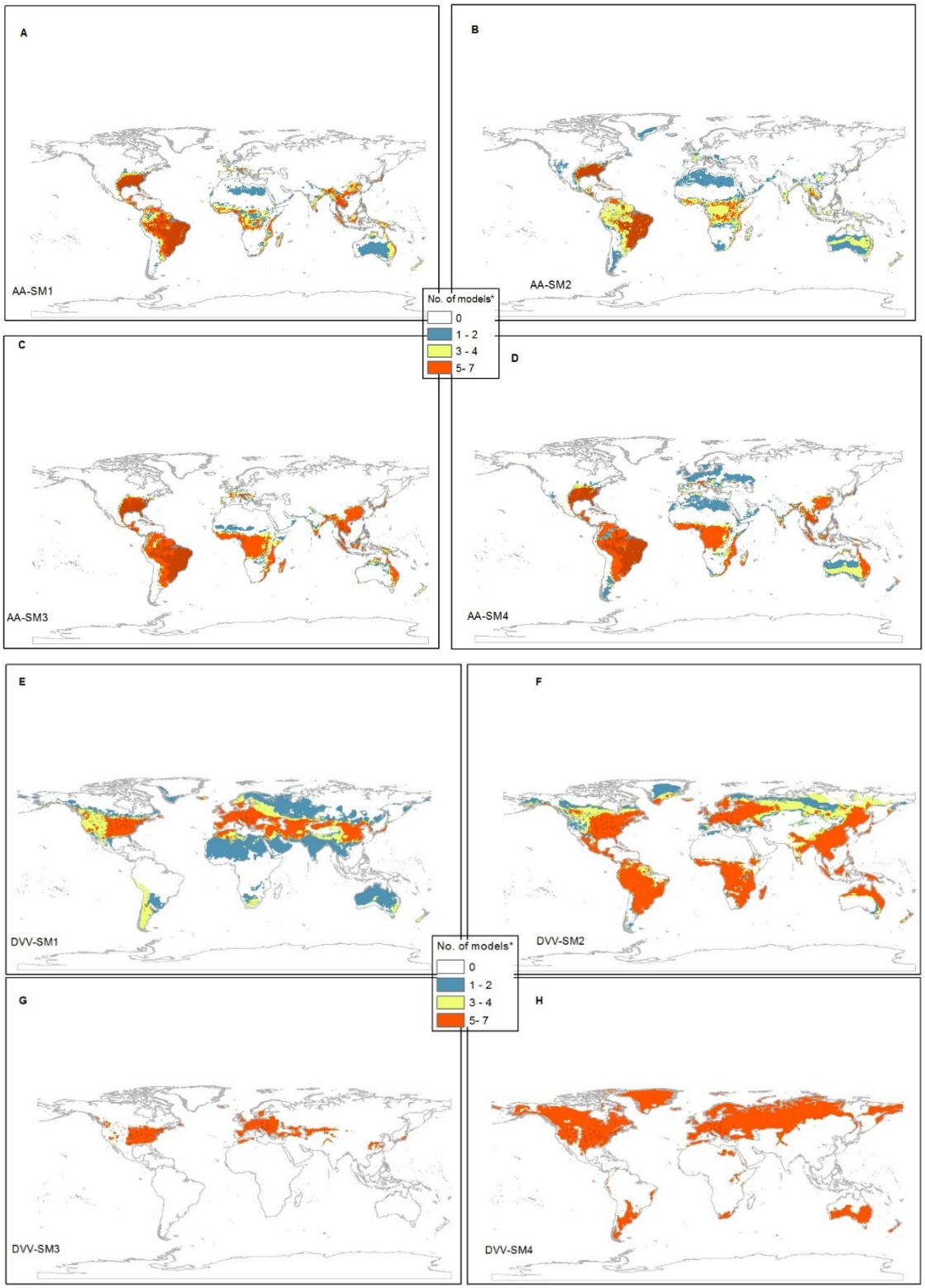


Above: Variable importance analysis for D. v. virgifera *background data delimitation. Graph labels show the coefficients of the three most important variables for* D. v. virgifera *over the given distances from presence points*



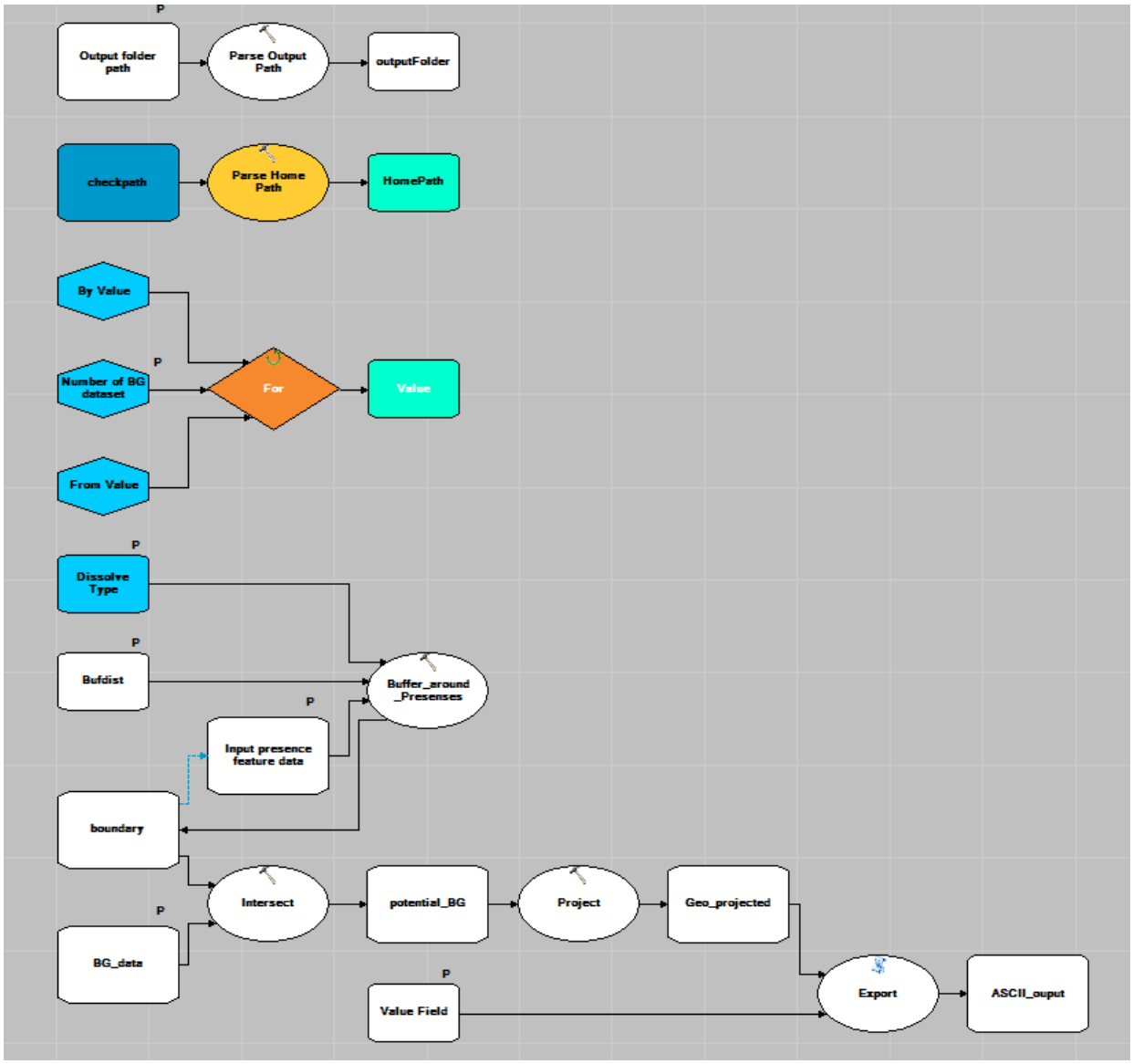
Left: Pseudo-absence points from the four pseudo-absence selection methods for D. v. virgifera *data. Pseudo-absence points plotted with presence points on the first three principal components of the training dataset (A) SM1, (B) SM2, (C) SM3, and (D) SM4*

Appendix 3.2 Model consensus maps based on different pseudo-absence methods



Sub -maps A-D show model agreement on predictions for A. albopictus *according to the four pseudo-absence selection methods (Sub -maps E-H correspond to* D. v. virgifera *predictions)*

Appendix 4.1 Automated framework to detect change in variable importance over distance



Description: Extracts background datasets around presence points according to the specified sequence of distances. Internally convert coordinate systems to enable converting projected coordinate systems into geographic coordinate systems after the background is extracted using plane circular buffers. Finally, exports extracted buffers using database files [txt, csv, ascii, dbf, info] for further PCA analysis using statistical software.

Appendix 4.2 Slope, Aspect and Hillshade data derivation

Slope is an inclination of or deviation of a surface from a horizontal or vertical line. It can be an important eco-geographical factor specially for characterizing access to a certain habitat both for plants and animals (Richards-Zawacki, 2009).

Slope dataset: slope is calculated by finding the ratio of the vertical (rise or descent) change divided by the horizontal (run) change between any two points on the terrain. Slope can be calculated by the equation given below.

$$Slope(deg) = \frac{rise}{run} = tan\theta \quad (Eq.\,3.1)$$

A GIS layer of slope can be calculated from a digital elevation model (DEM) using the equation 3.1 using spatial data analysis. Spatial modelling allows new attribute data derivation at any spatially explicit location by using a function of nearby pixels (using neighbourhood calculations) (Burrough & McDonell, 1998). In any given DEM the x, y and z values that correspond to the longitude, latitude and altitude of the dataset respectively can be used to generate slope using eq. 3.1. The slope is derived using the rates of change of the surface in the horizontal (dz/dx) and vertical (dz/dy) directions from the centre pixel. The ATAN function in ArcInfo which gives the inverse tangent of spatial grids is used to calculate the slope as follows the conversion rate 57.29578 is used to convert ATAN's radian calculations into degrees by using the 180/p conversion.

$$Slope_{degrees} = ATAN\left(\sqrt{(\left({}^{dz}/_{dx}\right)^2 + \left({}^{dz}/_{dx}\right)^2)}\right) * 57.29578 \quad (Eq.\,3.2)$$

Aspect is the horizontal orientation of any given slope on a terrain. Variation in aspect, for example the variation between north facing versus south facing slopes, was found to affect species composition of vegetation gradients along a given terrain (Astrom *et al.*, 2007). This variation is usually associated with the amount of sunlight the differently orientated slopes get where north facing slopes are moister than south facing slopes.

Aspect dataset: Aspect in a spatial dataset gives the downslope orientation of the maximum rate of change along each cell compared to its neighbouring pixels. The same representation

of x, y and z values of a spatial elevation data are used in calculating this layer. The value of an aspect dataset is given between 0-360. It is measured clockwise 0° degrees due north through Northeast (NE), East (E),Southeast (SE) etc… coming full circle to due north at 360°. A value of -1 signifies a flat surface.

$$aspect = 57.29578 * ATAN\left(\frac{dz/dy}{-(dz/dx)}\right) \quad (Eq.\ 3.3)$$

Where the pixel being processed will be 90.0 – aspect, if aspect <0; 360.0 – aspect + 90.0, if aspect > 90; and 90.0 – aspect, if aspect is between 0 and 90.

Hill shade is a geographic factor that can affect distribution of species, especially those that require a certain amount of shade or its absence to survive and/or thrive; for instance the highly invasive weed Purple loosestrife (*Lythrum salicaria*) is found to thrive in areas of little or no shade (Weihe & Neely, 1997).

Hillshade dataset: Hill shade is calculated in a GIS environment by simulating illumination from the sun over a terrain and map which part of the terrain is shaded and which part gets illuminated. This layer is calculated from a DEM. High values in this layer represent eastern slopes with high exposure to the sun while lower values represent shaded areas of western slopes. Two parameters are used in calculating the hillshade; these are the illumination angle and the illumination direction. Illumination angle is calculated by changing the altitude into a zenith angle (see equation 3.4) and illumination direction is calculated by converting the azimuth angle from geographic to mathematic angle (se equation 3.5).

$$Zenith_{rad} = (90 - Altitude) * \pi/180.0 \quad (Eq.\ 3.4)$$

Where altitude refers to the illumination source which is given in degrees above horizontal orientation, the altitude used in this study is 45°.

$$Azimuth_{rad} = (360.0 - Azimuth + 90.0) * \pi/180.0 \quad (Eq.\ 3.5)$$

If azimuth < 360.0 and

$$Azimuth_{rad} = ((360.0 - Azimuth + 90.0) - 360.0) * \pi/180.0 \quad (Eq.\ 3.6)$$

If azimuth ≥ 360.0, the azimuth used in this study was 315° therefore equation 3.5 applied.

The slope and aspect values needed to calculate the hillshade along with the zenith and azimuth of the illumination source are derived from equation 3.2 and .3.3 without the conversion rate that converts the slope and aspect values into degrees in order to calculate the hillshade.

$$Slope_{rad} = ATAN\left(Zfactor * \sqrt{\left(\left({}^{dz}/_{dx}\right)^2 + \left({}^{dz}/_{dx}\right)^2\right)}\right) \quad (Eq.\ 3.7)$$

$$Aspect_{rad} = ATAN\left({}^{dz}/_{dy} / \left(-{}^{dz}/_{dx}\right)\right) \quad (Eq.\ 3.8)$$

Where aspect in radians is defined in the range of 0-2p with 0 orienting towards the east and the following conditions are given, if dz/dx in equation 3.8 is non-zero and $aspect_{(rad)} < 0$ then the resulting $aspect_{(rad)} = 2 * p + aspect_{(rad)}$; if dz/dx=0 and dz/dy > 0 then $aspect_{(rad)} < p/2$; if dz/dx=0 and dz/dy <0 then $aspect_{(rad)} < 2 * p - p/2$; in all other cases $aspect_{(rad)}$ takes the value equation 3.8 solves to.

$$Hillshade = 255 * \left(\left(\cos(Zenith_{rad}) * \cos(Slope_{rad}) + \left(\sin(Zenith_{rad}) * \sin(Slope_{rad}) * \cos(Azimuth_{rad} - Aspect_{rad})\right)\right)\right) (Eq.\ 3.9)$$

The hillshade values are between 0-255 where a value < 0 is set to 0. The higher the value of a hillshade grid the darker it gets due to shading effect due to the elevation, slope and aspect conditions of the pixels in the model as well as its relative location from the source of illumination in this case the sun.

Appendix 4.3 Background on h-NLPCA dimension reduction method

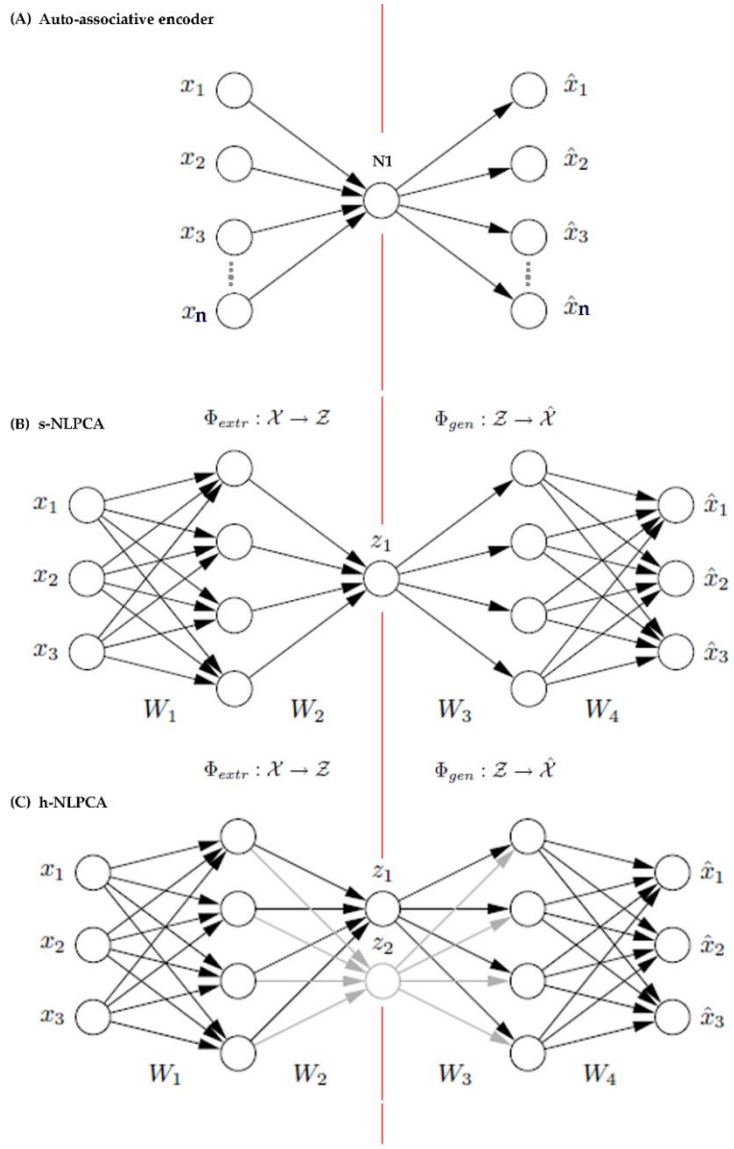
The description of the h-NLPCA by Scholz and Vigario (2002) is adapted here to give a brief background. The h-NLPCA (Figure 7.4.3-C) is built upon a pre-existing form of a multilayer perceptron (Bishop, 1995) network (sub figure A, next page) with an auto-associative neural network topology also known as bottleneck or hour glass topology (Marivate *et al.*, 2007).

The auto-associative network is a linear multilayer perceptron auto-encoder that has the same number of inputs (nodes) as outputs and have a hidden layer with fewer nodes. The weights in the network change while learning to minimize the mean square error. Due to this bottleneck architecture, the equivalent number of inputs and outputs, and the algorithm that minimizes the mean square error it was possible to converge the n features to the n^{th} dimension in the linear PCA feature space for a given $n \times m$ matrix (Baldi & Hornik, 1989).

Kramer (1991) then expanded the above auto-associative encoder into a non-linear PCA by adding two layers of nodes with non-linear functions at the start and end of the auto-associative encoder (sub figure B, next page). The extension enabled the linear auto-associative network to extract principal components from non-linear feature subspace.

In a nutshell, what Scholz and Vigario (2002) have done was extend the s-NLPCA into a hierarchical auto-associative neural network (non-linear PCA encoder) by superimposing an extra non-linear network with a function that applies hierarchy constraints to the feature space in the same way as the linear PCA does (Scholz & Vigario, 2002). This gave rise to the h-NLPCA (sub figure C, next page). The symmetrical NLPCA (s-NLPCA) has similarity to h-NLPCA in mapping data into a non-linear feature sub-space but it lacks the capability to discriminate features. Generally, an s-NLPCA (Figure 7.4.3-B) is sufficient if the problem involves only reducing dimensions and does not require feature selection (Scholz & Vigario, 2002; Gorban, 2007). However, in the context of this study a method that does dimension reduction as well as is capable of identifying features in the non-linear feature space is preferable as there is no subsequent feature selection step specified for datasets that are treated with dimension reduction, hence h-NLPCA was the appropriate choice because the

hierarchical learning algorithm allows for feature identification as well as dimension reduction.



Topologies of the linear auto-encoder (A)[34]*, the s-NLPCA (B) and the h-NLPCA (C) auto-associative neural networks. The illustrations for the topologies of (B) s-NLPCA [3-4-4-4-3]*[35]*and (C) h-NLPCA ([3-4-2-4-3] + [3-4-1-4-3]) networks were taken from Scholz et al. (2008, pp. 49,50). The illustration for (A) the auto-encoder [4-1-4] was adapted from Marivate et al. (2007, p. 2) Both s-NLPCA and h-NLPCA are extensions of the linear auto-encoder. In both cases the left side of the red line in the middle show the first part of the dimension reduction process where data are extracted non-linearly from the inputs in [x1, x2, x3 …..] and linearly decoded at [z_1, (z_2)] (function Φ_{extr}); the right side of the red line shows where data is linearly decoded from [z_1, (z_2)] and are non-linearly generated at the output [x1, x2, x3…] (function Φ_{gen}). For the h-NLPCA an additional of 3-4-1-4-3 topology network is transposed on top of the s-NLPCA topology so that learning error is separately computed 1) for the sub-network (E_1) and 2) on the sub-network + the whole network ($E_{1,2}$), and later added to produce the total hierarchic error ($E=E_1+E_{1,2}$) which is used to update the weights through the whole network. This hierarchic learning enables the h-NLPCA to do feature extractions as well as dimension reductions. Refer Scholz and Vigario (2002) and Scholz et al. (2008) for detail model specification the above description was also consulted from the same references.*

[34] The auto-encoder can have more than one nodes as long as the number of nodes at the hidden layer are fewer than the nodes at the input and the output which have equal number of nodes (Marivate *et al.*, 2007).

[35] Network topology description that gives the number of nodes in input, any hidden layers and output layers in that order.

Appendix 4.4 Ranks of variables as per proportions of their use in the tested models

No.	Variables	Rank	Proportion
1	Annual mean temperature (°C)	5	0.36
2	Mean diurnal temperature range (mean(period max-min)) (°C)	6	0.33
3	Isothermality (Bio02 ÷ Bio07)	16	0.04
4	Temperature seasonality (C of V)	9	0.27
5	Max temperature of warmest week (°C)	6	0.33
6	Min temperature of coldest week (°C)	8	0.29
7	Temperature annual range (Bio05-Bio06) (°C)	12	0.16
8	Mean temperature of wettest quarter (°C)	4	0.38
9	Mean temperature of driest quarter (°C)	10	0.24
10	Mean temperature of warmest quarter (°C)	8	0.29
11	Mean temperature of coldest quarter (°C)	8	0.29
12	Annual precipitation (mm)	1	0.67
13	Precipitation of wettest week (mm)	11	0.22
14	Precipitation of driest week (mm)	10	0.24
15	Precipitation seasonality (C of V)	10	0.24
16	Precipitation of wettest quarter (mm)	1	0.67
17	Precipitation of driest quarter (mm)	2	0.44
18	Precipitation of warmest quarter (mm)	3	0.40
19	Precipitation of coldest quarter (mm)	4	0.38
20	Annual mean radiation (W m^{-2})	13	0.11
21	Highest weekly radiation (W m^{-2})	12	0.16
22	Lowest weekly radiation (W m^{-2}	7	0.31
23	Radiation seasonality (C of V)	14	0.09
24	Radiation of wettest quarter (W m^{-2})	18	0.00
25	Radiation of driest quarter (W m^{-2})	12	0.16
26	Radiation of warmest quarter (W m^{-2})	8	0.29
27	Radiation of coldest quarter (W m^{-2})	14	0.09
28	Annual mean moisture index	15	0.07
29	Highest weekly moisture index	13	0.11
30	Lowest weekly moisture index	15	0.07
31	Moisture index seasonality (C of V)	15	0.07
32	Mean moisture index of wettest quarter	17	0.02
33	Mean moisture index of driest quarter	14	0.09
34	Mean moisture index of warmest quarter	18	0.00
35	Mean moisture index of coldest quarter	16	0.04
36	Elevation (m)	10	0.24
37	Slope (deg)	17	0.02
38	Aspect (deg)	18	0.00
39	Hillshade	15	0.07

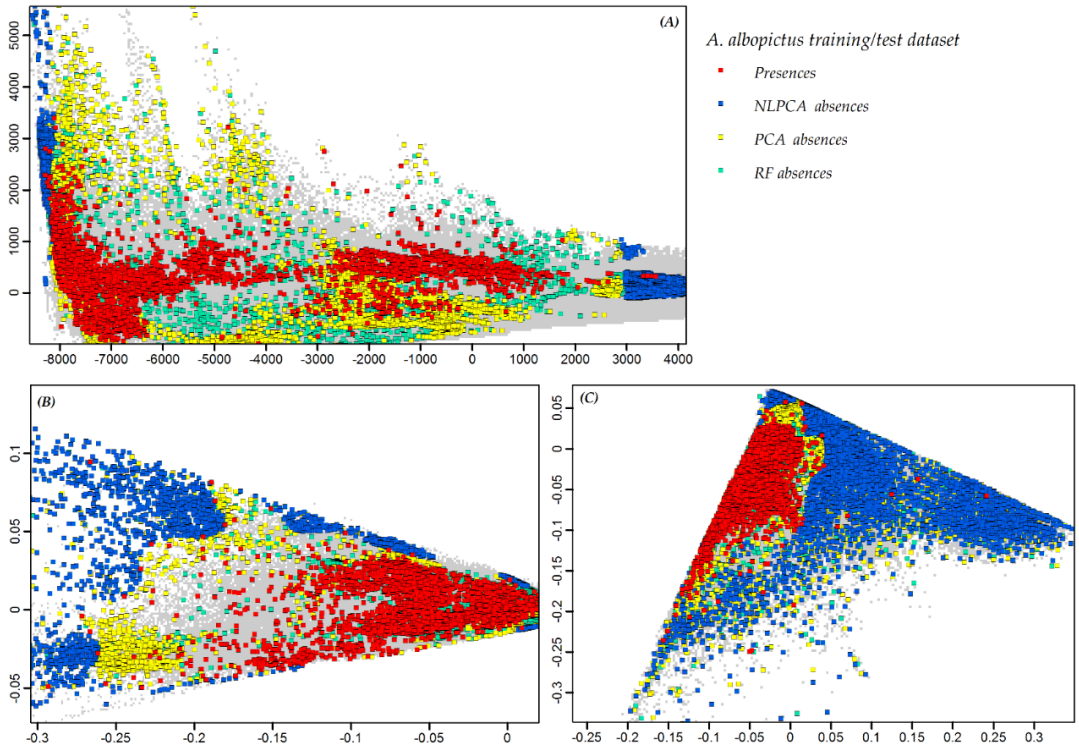
Appendix 4.5 Comparison of prediction accuracy for different species, dimension reduction, model result combinations

No.	Combination	means	
1	aa.dr3.svm	0.986	a
2	vv.dr3.svm	0.981	a
3	aa.dr2.svm	0.980	ab
4	aa.dr3.cart	0.977	ab
5	vv.dr3.cart	0.969	abc
6	vv.dr2.svm	0.964	abc
7	dvv.dr3.svm	0.954	abc
8	aa.dr2.cart	0.947	abc
9	dvv.dr2.svm	0.943	abc
10	aa.dr1.svm	0.928	abc
11	vv.dr1.svm	0.924	abc
12	vv.dr2.cart	0.917	abc
13	ag.dr3.svm	0.914	abc
14	vv.dr3.qda	0.914	abc
15	aa.dr1.cart	0.903	abc
16	aa.dr3.qda	0.900	abc
17	dvv.dr3.cart	0.887	abcd
18	vv.dr2.qda	0.880	abcd
19	aa.dr1.qda	0.876	abcd
20	vv.dr1.cart	0.871	abcd
21	vv.dr1.qda	0.867	abcd
22	ag.dr3.qda	0.837	abcd
23	ag.dr2.svm	0.835	abcd
24	aa.dr2.qda	0.830	abcd
25	ag.dr3.cart	0.827	abcd
26	vv.dr1.log	0.825	abcd
27	ag.dr2.qda	0.808	abcde
28	dvv.dr3.qda	0.789	abcdef
29	ag.dr1.svm	0.788	abcdef
30	aa.dr3.log	0.788	abcdef
31	dvv.dr2.cart	0.784	abcdef
32	vv.dr3.log	0.782	abcdef
33	dvv.dr3.log	0.779	abcdef
34	vv.dr2.log	0.777	abcdef
35	ag.dr2.cart	0.777	abcdef
36	ag.dr1.log	0.772	abcdef
37	ag.dr1.qda	0.765	abcdef
38	tp.dr3.svm	0.764	abcdef
39	ag.dr3.log	0.755	abcdef
40	ag.dr1.cart	0.736	abcdef

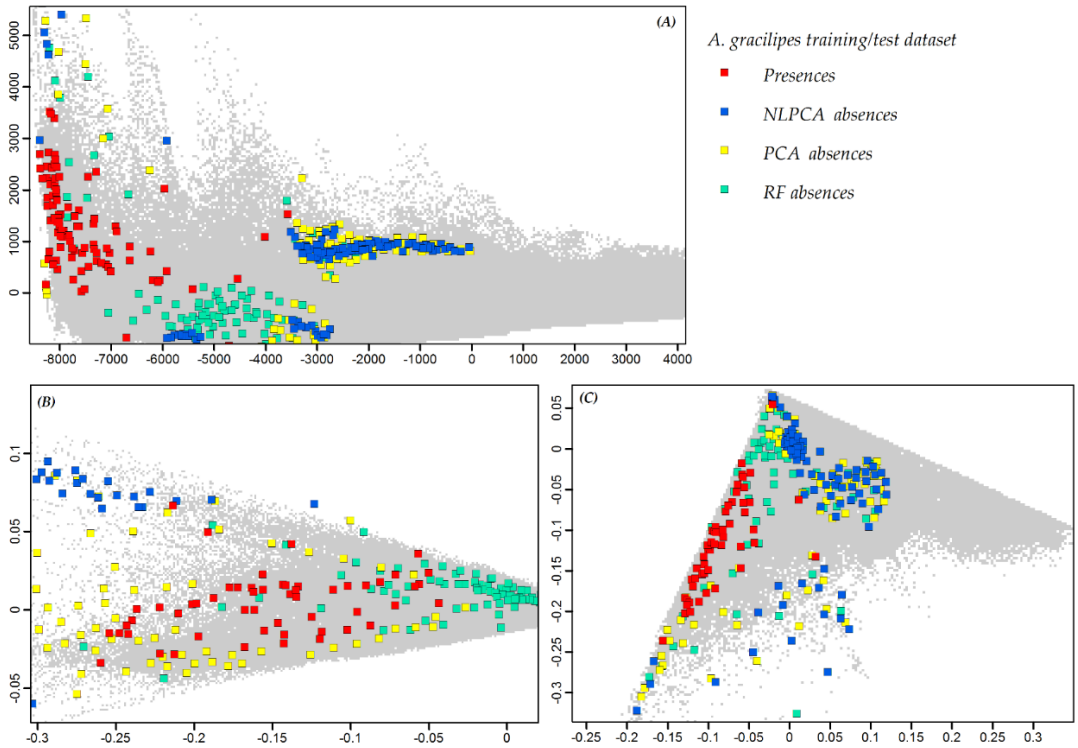
No.	Combination	means	
41	ag.dr2.log	0.733	abcdef
42	aa.dr1.log	0.723	abcdef
43	tp.dr1.svm	0.699	abcdef
44	dvv.dr1.svm	0.683	abcdef
45	tp.dr3.cart	0.674	abcdef
46	tp.dr2.svm	0.673	abcdef
47	tp.dr2.cart	0.665	abcdef
48	dvv.dr2.qda	0.663	abcdef
49	tp.dr1.qda	0.652	abcdef
50	dvv.dr1.cart	0.622	abcdef
51	tp.dr1.cart	0.613	abcdef
52	tp.dr3.qda	0.600	abcdef
53	tp.dr2.qda	0.589	abcdef
54	dvv.dr1.qda	0.569	abcdef
55	dvv.dr2.log	0.526	bcdef
56	dvv.dr1.log	0.517	cdef
57	tp.dr1.log	0.436	def
58	tp.dr2.log	0.367	ef
59	tp.dr3.log	0.350	f
60	aa.dr2.log	0.338	f

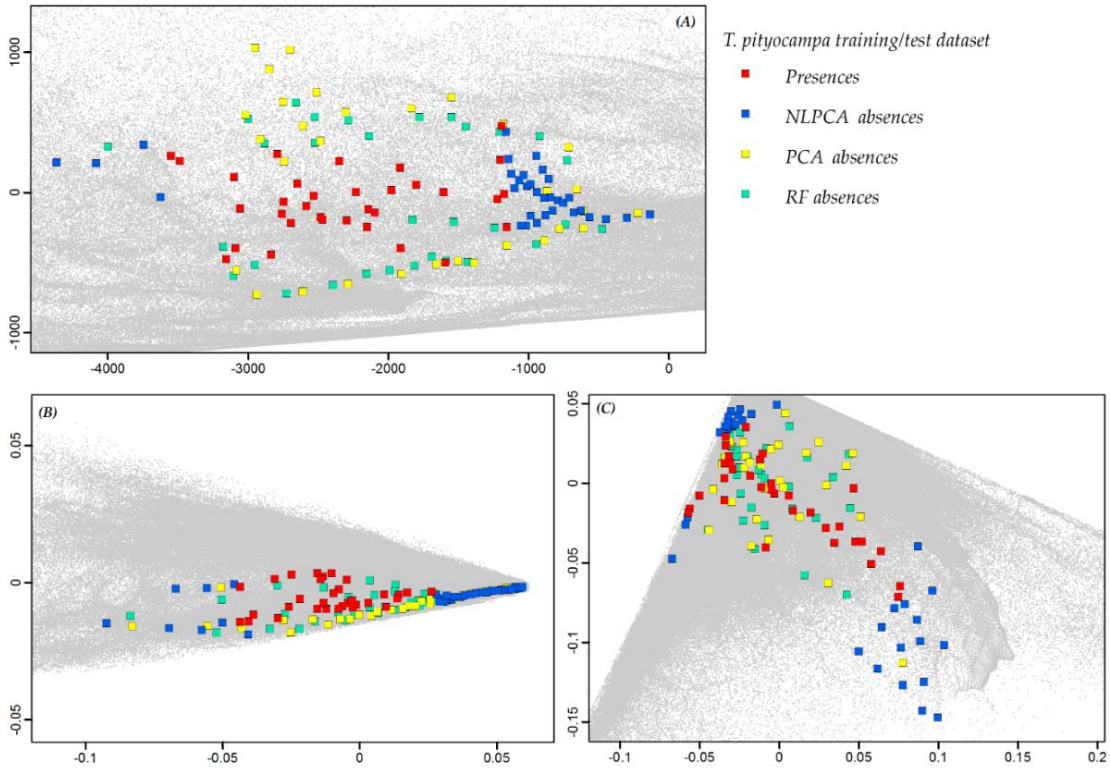
Variation in model mean Kappa scores according to different species data (SP), dimension reduction methods (DR) and model types (MT) combinations. Bars with different letters are significantly different (Tukey's HSD test, HSD = 0.45, $\alpha = 0.05$).

Appendix 4.6 Presence and pseudo-absence points in the environmental space

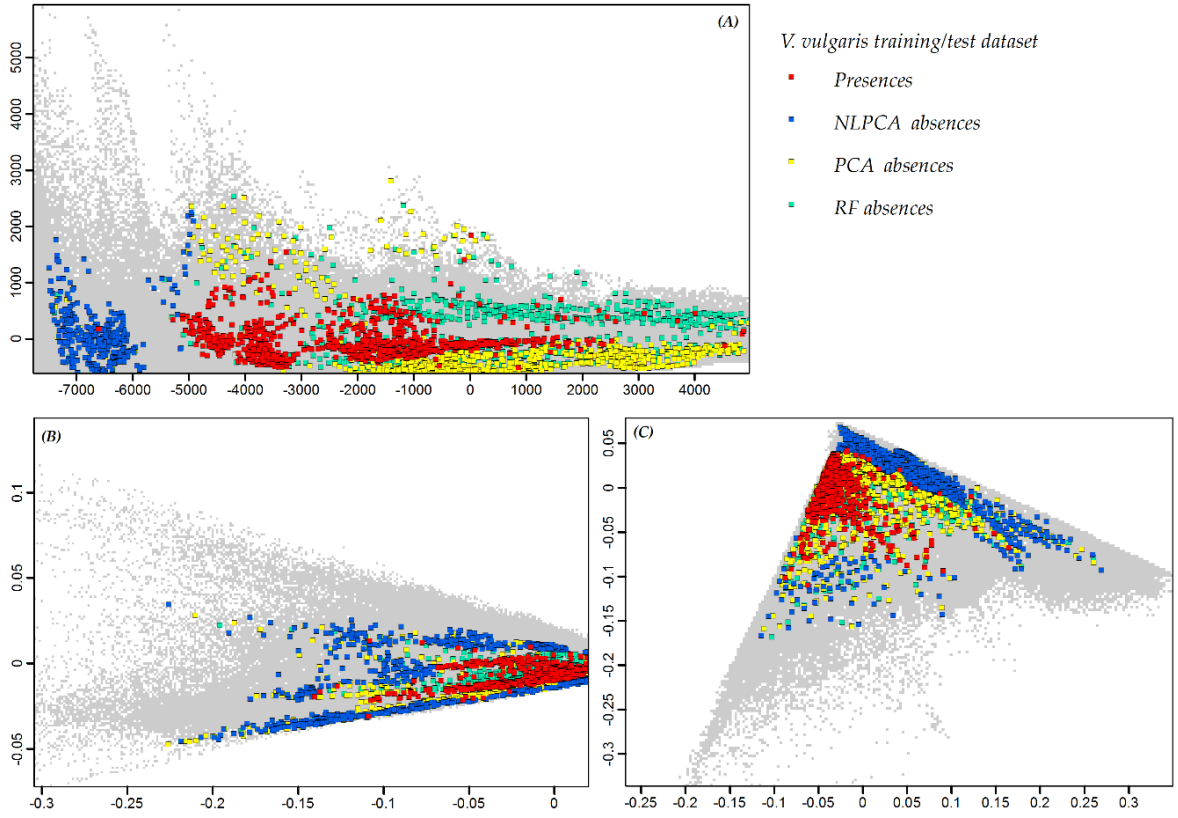


Presence and pseudo-absence points for A. albopictus *(above) and* A. gracilipes *(below)*

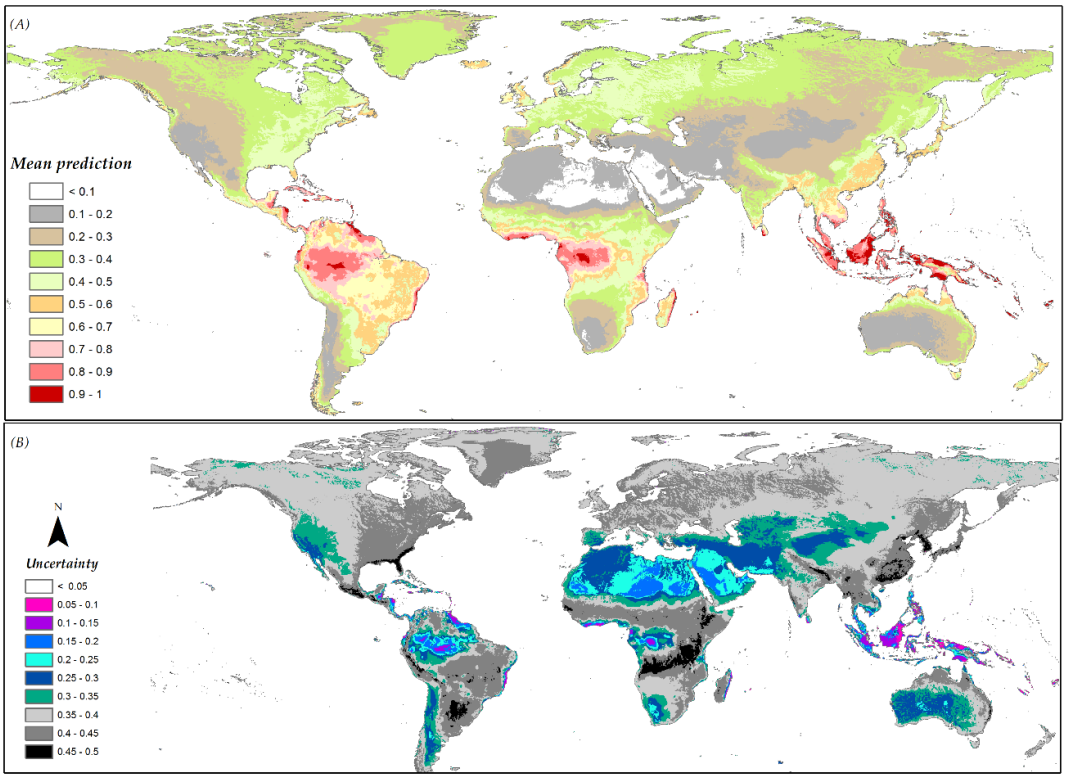




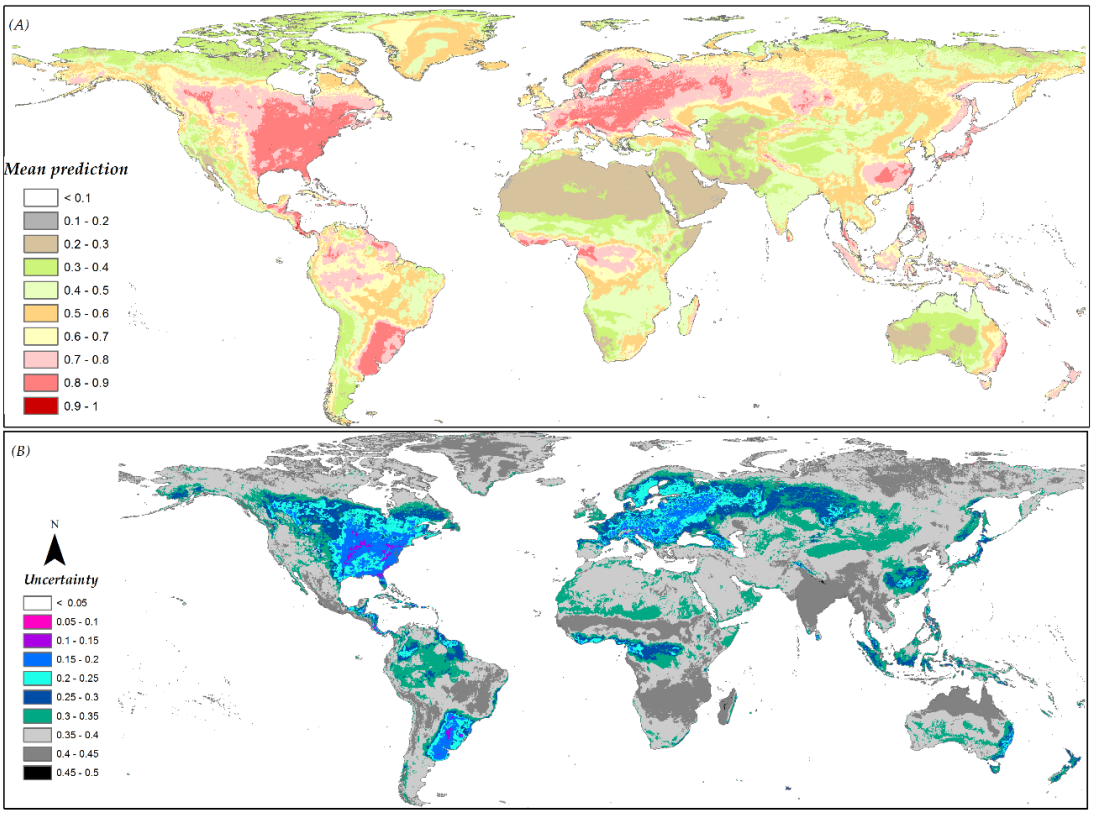
Presence and pseudo-absence points for T. pityocampa *(above) and* V. vulgaris *(below)*

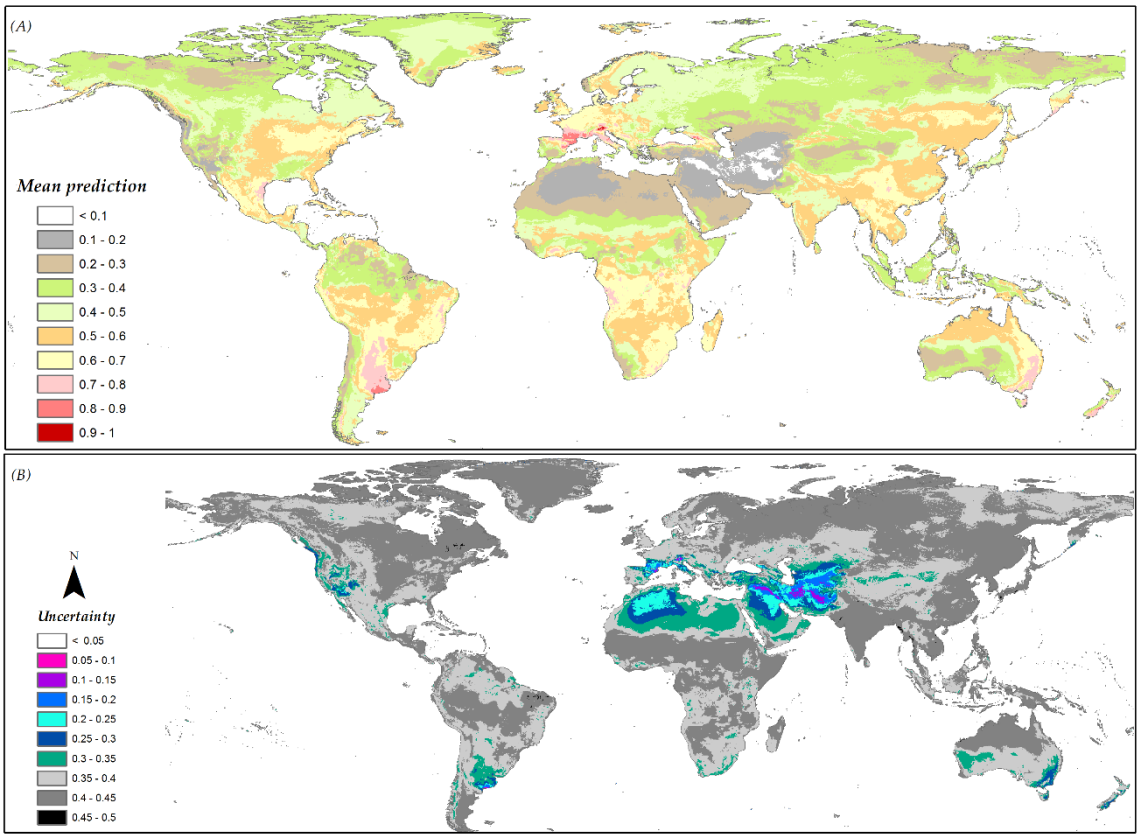


Appendix 4.7 Ensemble mean predictions and uncertainty maps

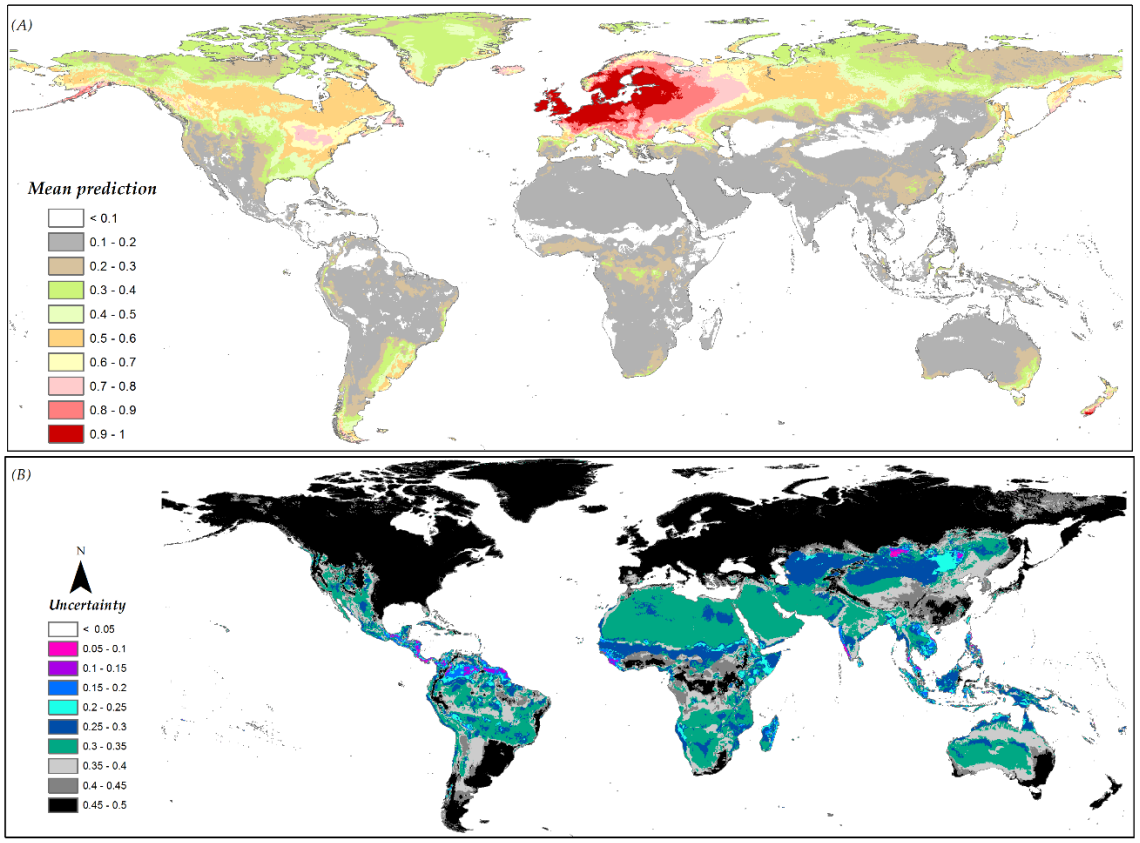


Average prediction (A) and uncertainty map (B) for A. gracilipes *(above) and* D. v. virgifera *(below)*

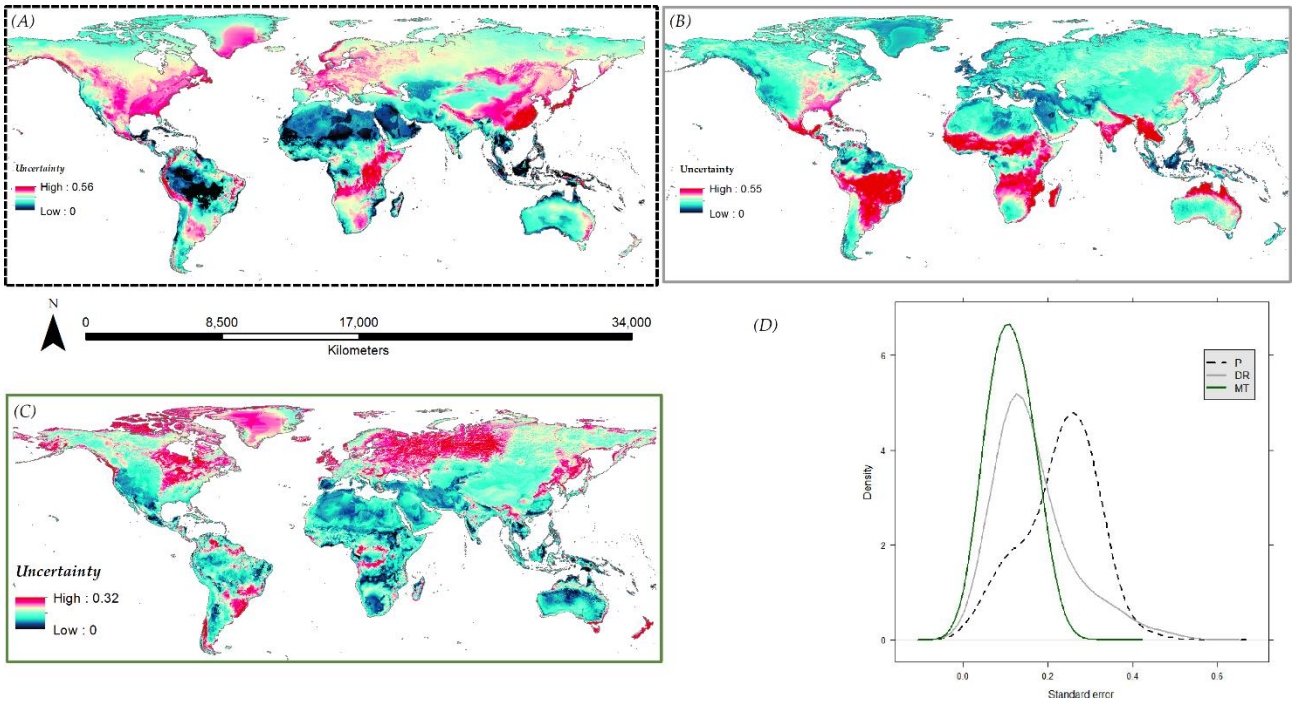




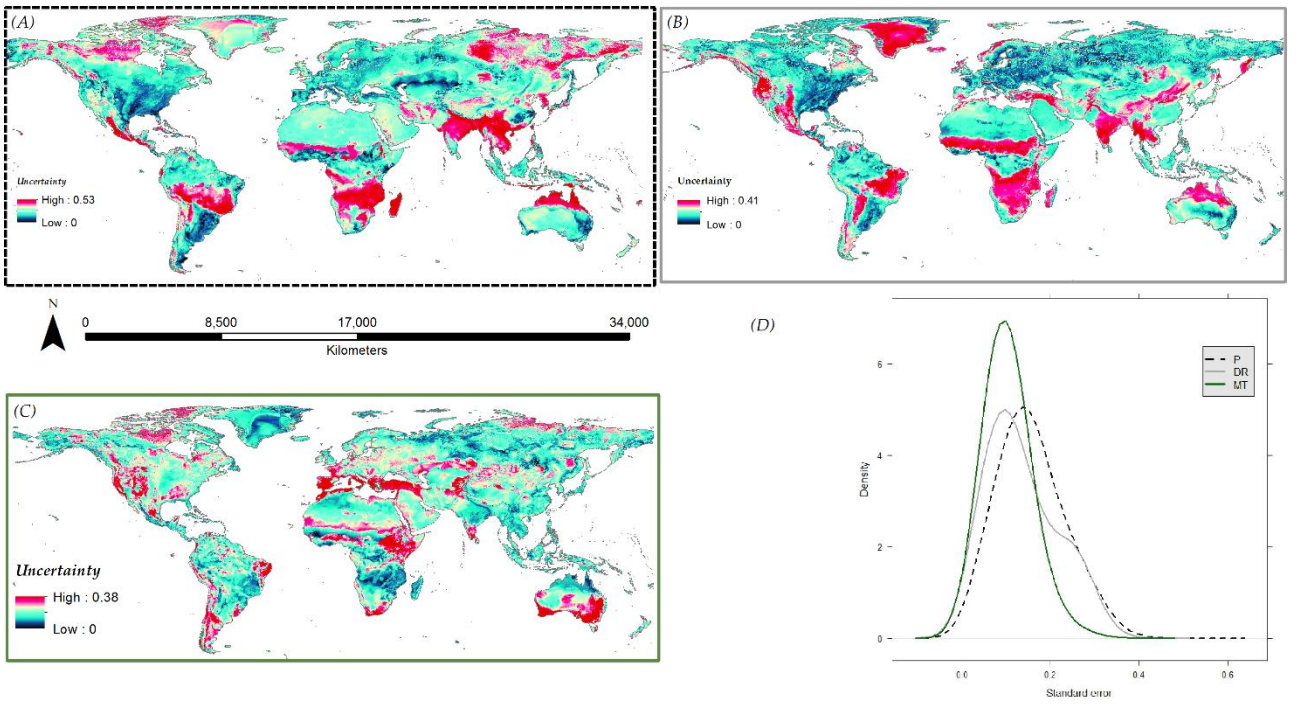
Average prediction (A) and uncertainty map (B) for T. pityocampa *(above) and* V. vulgaris *(below)*

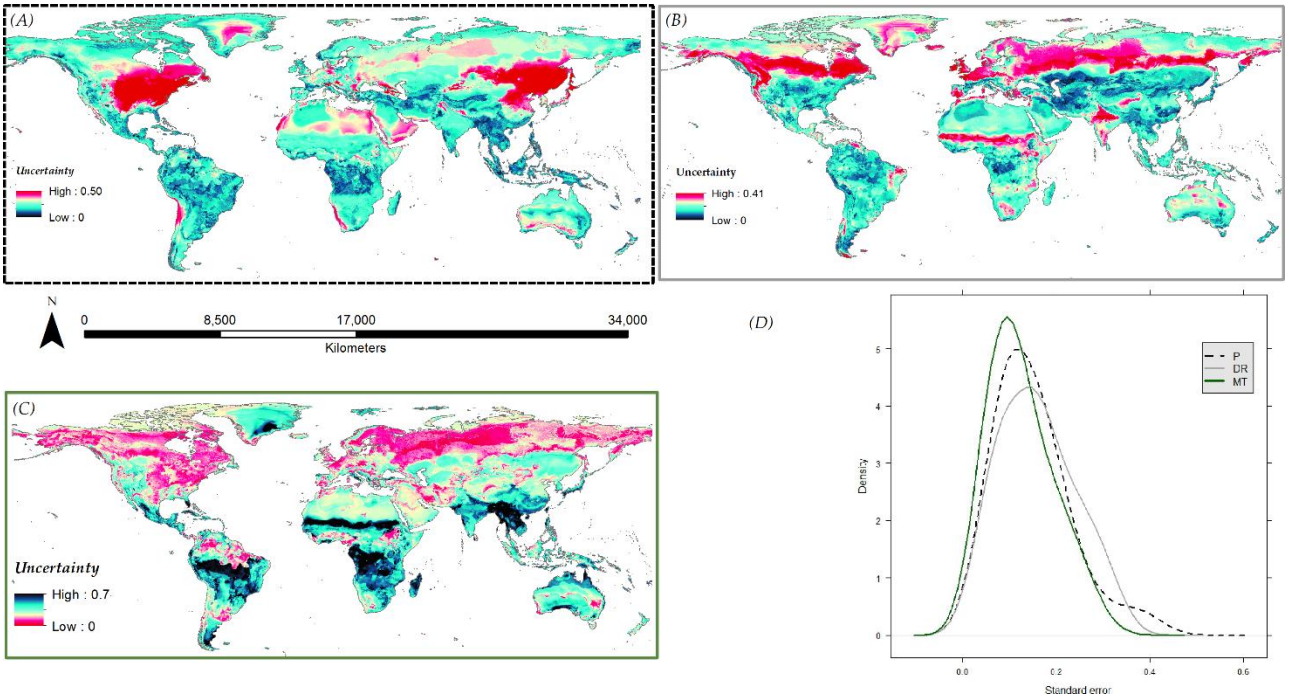


Appendix 4.8 Map of uncertainty by modelling components

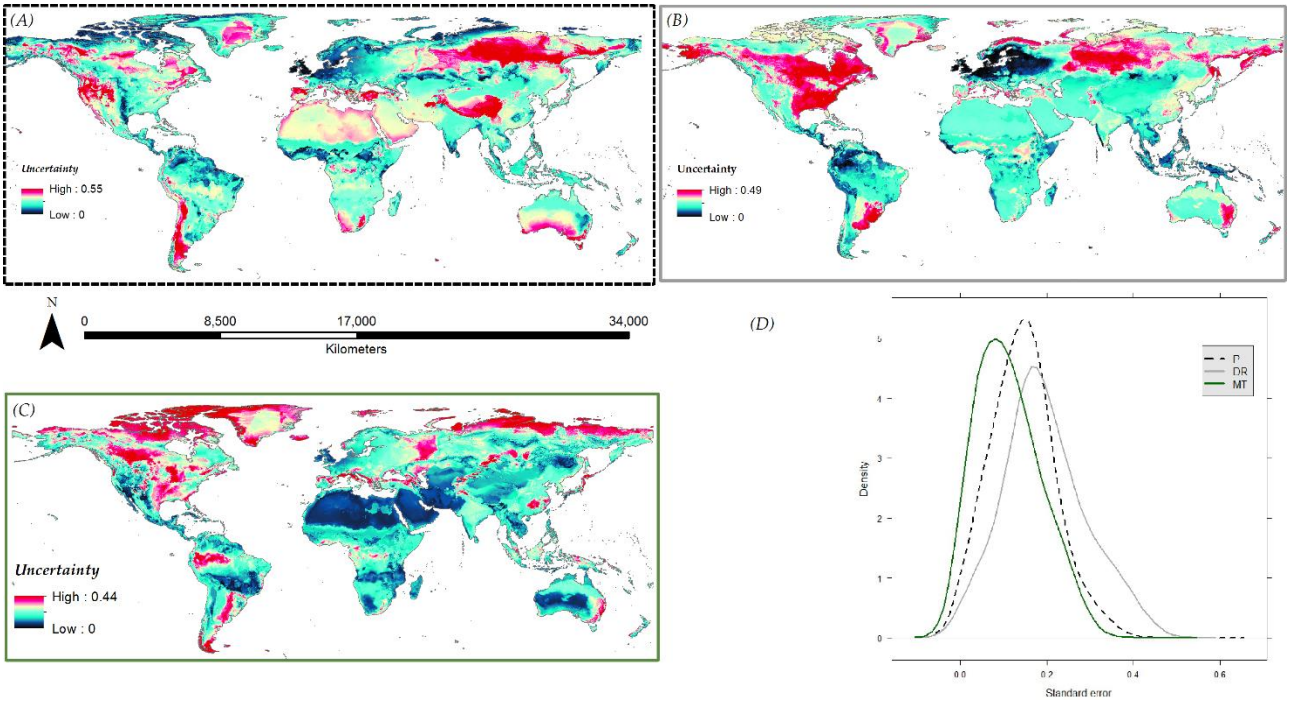


Standard error of predictions according to predictor data (A) model type (B) and dimension reduction (C) for A. gracilipes *(above) and* D. v. virgifera *(below)*

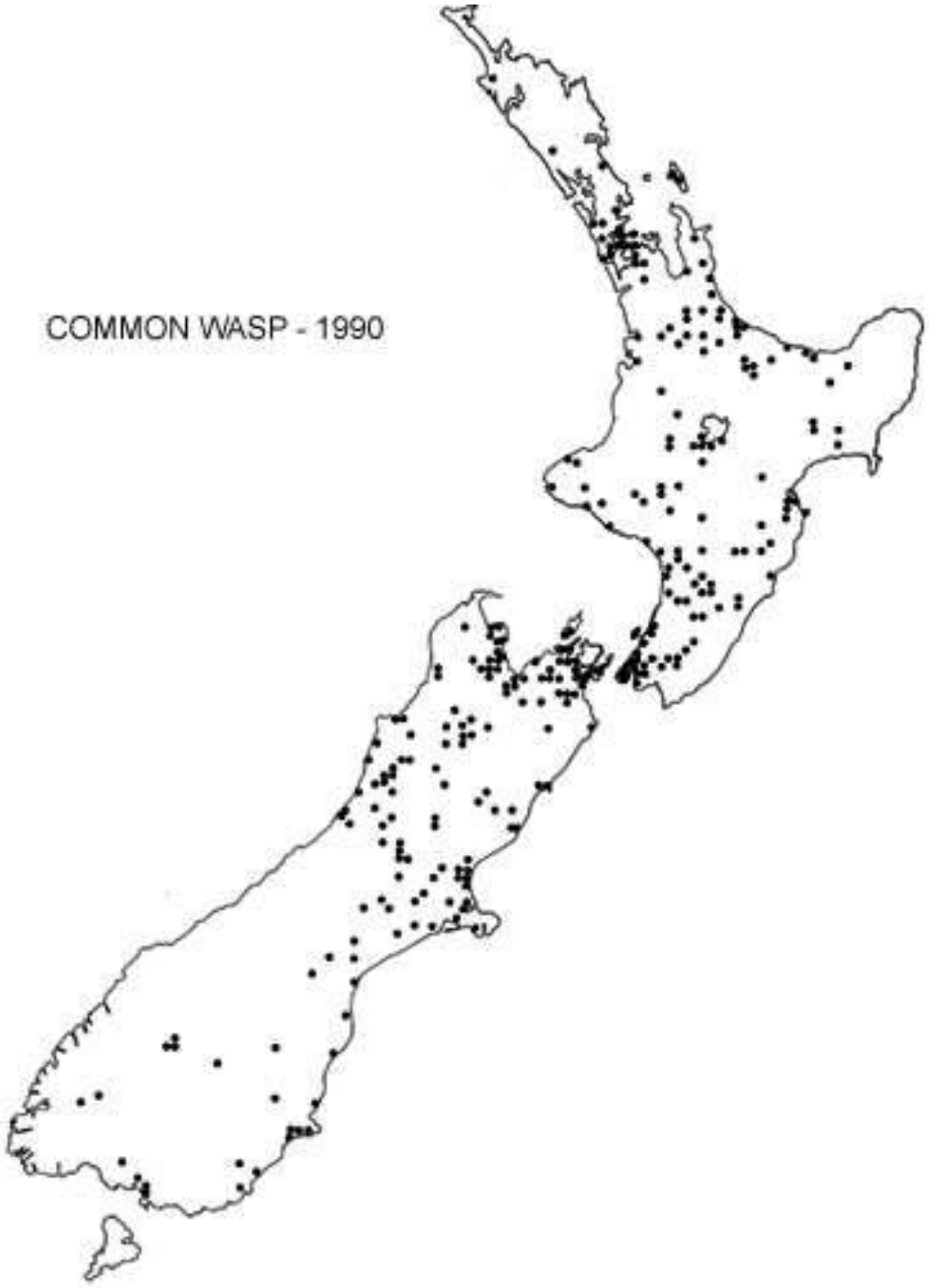




Standard error of predictions according to predictor data (A) model type (B) and dimension reduction (C) for T. pityocampa *(above) and* V. vulgaris *(below)*



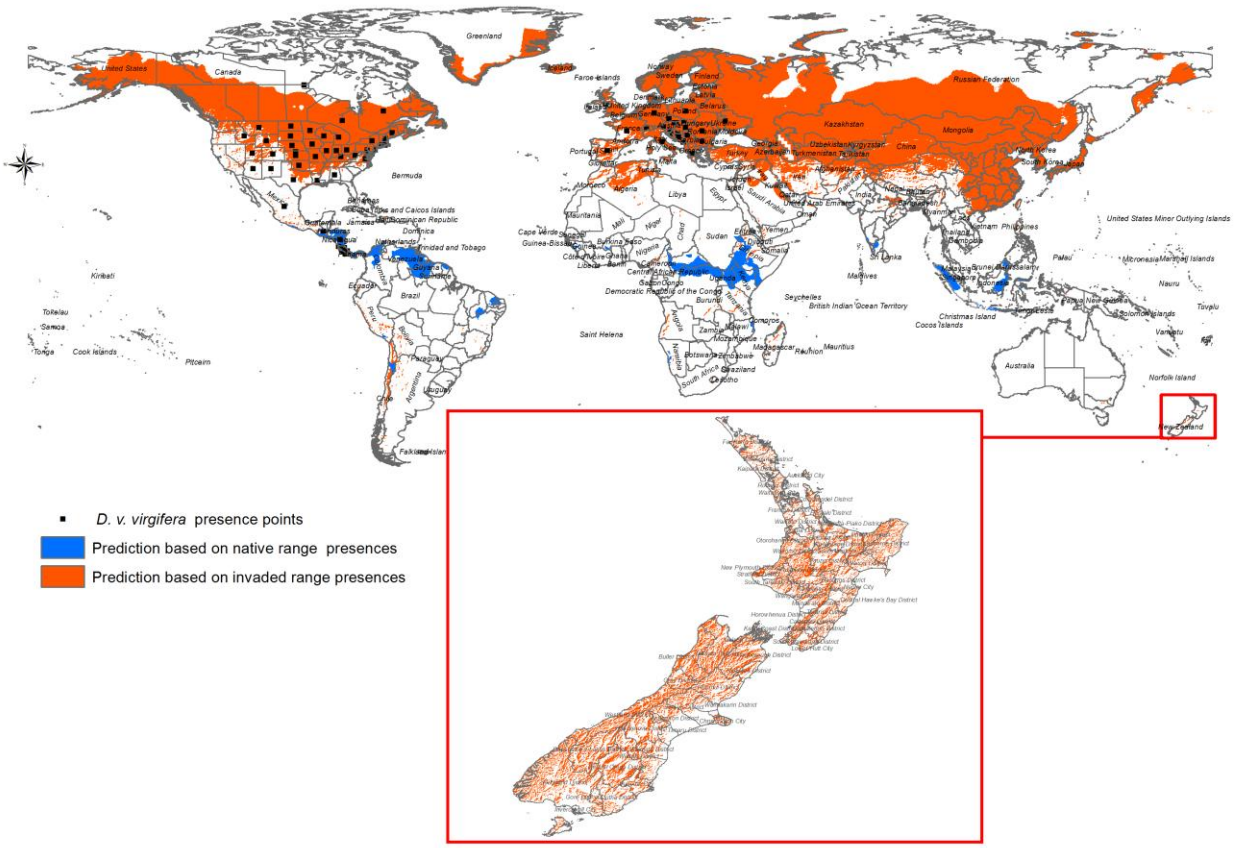
Appendix 4.9 External validation data for *V. vulgaris* in New Zealand



Geographic locations of V. vulgaris *presences in New Zealand. Source: AgResearch Research Centre, New Zealand.*

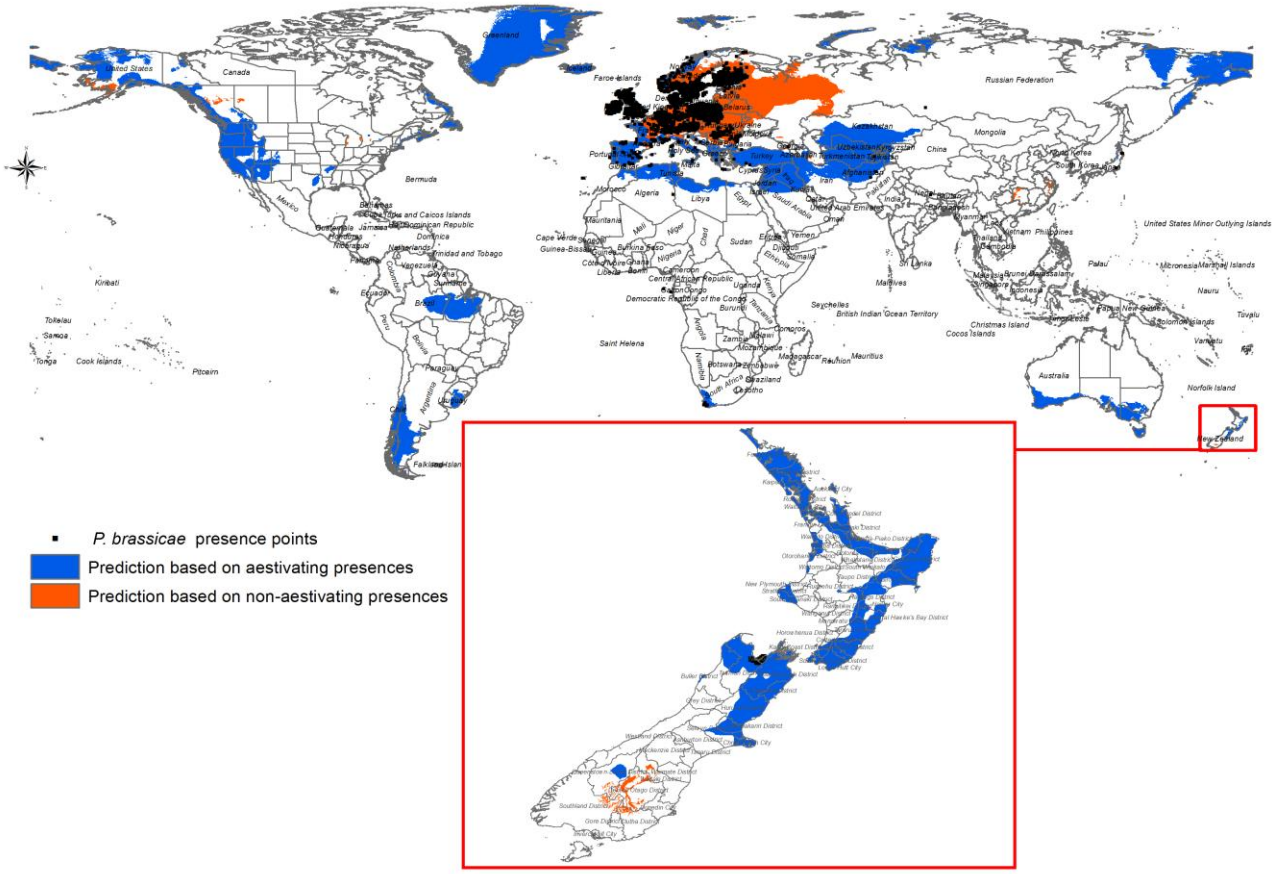
Appendix 5.1 Individual component model predictions based on different clusters of presence points.

1. *D. v. virgifera* component predictions



The individual prediction from the native range presences identified limited areas with the Central American range identified uniquely by the native range presence predictions. While most of the prediction from the native range scenario was also predicted by the model trained on the invaded range presences. More areas in East and Central Africa were predicted using the native range presence model. It is possible to notice on the global map that there were no native or invaded range based predictions for Australia and only limited predictions for New Zealand. However on the high resolution (30') prediction for New Zealand large areas were predicted in conformation with the other studies on potential distribution of D. v. virgifera (Aragón et al., 2010; Dupin et al., 2011; Senay et al., 2013). Regarding the prediction in Australia, it might be the case that the combined prediction led to the under estimation of the potential prediction in the Australian continent even though it accurately described the native Central American range as well as all the other areas in N. America and Europe and also identified more areas in Central and East Africa. Hence, it might be important to consider the choice of modelling with split predictions if the target areas were in Central and East Africa or Central America, otherwise go with the usual direct prediction methods for other study areas.

2. *P. brassicae* component predictions



Note that the global prediction in New Zealand showed less areas as suitable when compared to the high resolution prediction for the New Zealand extent. This shows the effect of scale on prediction outcomes and should be considered in comparing predictions from different studies (Austin & Van Niel, 2011).

Appendix 6.1 Rescaling the potential niche surface into a physiological suitability surface

```
# ---------------------------------------------------------------------------
# RescaleMechModel.py
# Created on: 2013-10-29 by Senait D. Senay
#  Description: Rescales a fundamental thermal (or any other scenopoetic variable) niche surface
#between 0 and 1 provided that an optimum temperature value is given
# ---------------------------------------------------------------------------
# Import arcpy module
import arcpy
# Check out any necessary licenses
arcpy.CheckOutExtension("spatial")
# Script arguments
mov = arcpy.GetParameterAsText(0)
if mov == '#' or not mov:
    mov = "15" # provide an optimum temperature- (mov) most optimum value

PhySuitability = arcpy.GetParameterAsText(1)
if PhySuitability == '#' or notPhySuitability:
    PhySuitability = "Path\to\output" # provide output filename
# Local variables:
MechInput = "GlobAvgTemp"
globniche1 = mov
SN1 = globniche1
min_gn1 = SN1
rescaled1 = min_gn1
GN1zone = SN1
max_gn1 = GN1zone
globniche2 = mov
globnich2r = globniche2
SN2 = globnich2r
GN2zone = SN2
min_gn2 = GN2zone
rescaled2 = min_gn2
max_gn2 = GN2zone
reverser = "path\to\output" # output file name for a -1 file to reverse assort the grid for val > mov
# 1 SplitRaster
arcpy.gp.RasterCalculator_sa("Con(( \"%MechInput%\"<  %mov%), \"%MechInput%\" , -9999)", globniche1)
# 2 DiscardNoDataValues
arcpy.gp.RasterCalculator_sa("Pick(\"%globniche1%\" != -9999,\"%globniche1%\")", SN1)
# 3 CreateZone for ZoneStatistics
arcpy.gp.CreateConstantRaster_sa(GN1zone, "1", "INTEGER", SN1, SN1)
# 4 create a MinimumRaster value
arcpy.gp.ZonalStatistics_sa(GN1zone, "VALUE", SN1, min_gn1, "MINIMUM", "DATA")
# 5 create a MaximumRaster value
arcpy.gp.ZonalStatistics_sa(GN1zone, "VALUE", SN1, max_gn1, "MAXIMUM", "DATA")
# 6 rescale the first half of the raster
arcpy.gp.RasterCalculator_sa("(\"%SN1%\"-\"%min_gn1%\")/Float(\"%max_gn1%\"-\"%min_gn1%\")",
rescaled1)
# 7 SplitRaster 2
arcpy.gp.RasterCalculator_sa("Con((\"%MechInput%\"  > %mov%) ,\"%MechInput%\" , -9999)", globniche2)
# 8 create a -1 value raster to reverse assort the raster with < optimum temp. values
arcpy.gp.CreateConstantRaster_sa(reverser, "-1", "INTEGER", "0.166666675359011", "GlobAvgTemp")
# 9 multiply the reverser raster with the raster < Mov values
arcpy.gp.Times_sa(globniche2, reverser, globnich2r)
# 10 DiscardNoDataValues 2
arcpy.gp.RasterCalculator_sa("Pick(\"%globnich2r%\" != 9999,\"%globnich2r%\")", SN2)
# 11 CreateZone for ZoneStatistics 2
arcpy.gp.CreateConstantRaster_sa(GN2zone, "1", "INTEGER", SN2, SN2)
# 12 create a MinimumRaster value 2
arcpy.gp.ZonalStatistics_sa(GN2zone, "VALUE", SN2, min_gn2, "MINIMUM", "DATA")
# 13 create a MaximumRaster value 2
arcpy.gp.ZonalStatistics_sa(GN2zone, "VALUE", SN2, max_gn2, "MAXIMUM", "NODATA")
# 14 rescale the second half of the raster 2
arcpy.gp.RasterCalculator_sa("(\"%SN2%\"-\"%min_gn2%\")/Float(\"%max_gn2%\"-\"%min_gn2%\")",
rescaled2)
# 15 Combine the two rescaled rasters
arcpy.gp.RasterCalculator_sa("Con((\"%rescaled1%\"!=-9999),\"%rescaled1%\",\"%rescaled2%\")",
PhySuitability)
```

Appendix 7.1 Data extracted from the Atlas of the Insects of The British Isles

INTRODUCTION

This Provisional Atlas of the Insects of the British Isles has been prepared using the methods evolved by Dr. Franklyn Perring for the "Atlas of the British Flora" published for the Botanical Society of the British Isles by Thomas Nelson and Sons Ltd. in 1962.

The maps in Part 1 have been prepared from records sent in to the Biological Records Centre by participants in the Lepidoptera Distribution Maps Scheme with the addition of some data from the literature. Records received up to 31st December, 1969 are included. Up-dated maps will be prepared from time to time as the scheme progresses.

Each spot represents a verified record from the 10 Km square in which it appears.

Only two date classes are shown on the maps although records are being collected from three periods. For the present series of maps the two date classes 'pre-1940' and '1940 to 1960' have been lumped together. When more journals have been scrutinised it is hoped that sufficient data will become available to enable these two periods to be separated.

Only the resident British species are included.

The univoltine and bivoltine forms of Aricia agestis (Brown Argus and Northern Brown Argus) are treated as separate species although some authors consider these to be only sub-species.

A situation map is given which shows from which squares records have been received. By comparing the maps of species with this map it is possible to determine to some extent whether or not the absence of a species from any particular area is real, or likely to be due to absence of records. Later situation maps will indicate the actual number of species recorded in each square and will thus enable more precise deductions to be made.

A composite map of Rhamnus cathartica and Frangula alnus, the food plants of Gonepteryx rhamni (Brimstone) have been included as there is already a very good correlation of this species with its food plant.

Similarly good correlations of the calcicole species, Lysandra coridon (Chalk Hill Blue) and Lysandra bellargus (Adonis Blue) are already shown with the appropriate geological features.

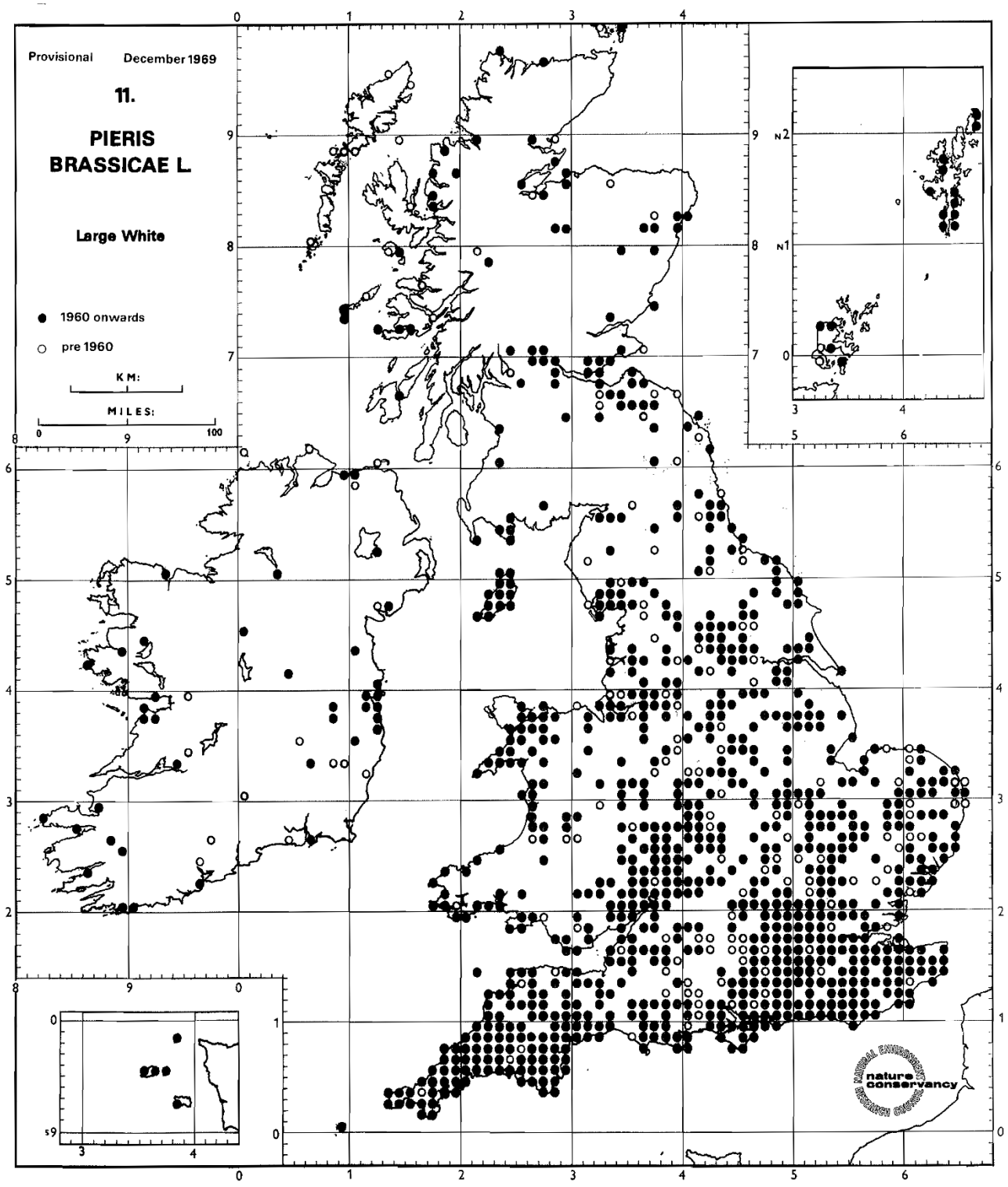
Although other correlations may be apparent great care should be taken in interpreting the data especially the ratio of 'pre-1960' to '1960 onwards' records. In some cases absence of records in the '1960 onwards' class will be due to incomplete data rather than actual absence.

Acknowledgments

To all the 620 recorders who contributed records used for the maps our best thanks are due. They will be acknowledged personally in the final Atlas. I should also like to thank Dr. F.H. Perring for his support and encouragement and the assistant staff of the Biological Records Centre, especially Rosemarie Cooke, Elizabeth Tasker and Michael Skelton who were responsible for processing the data and making the maps.

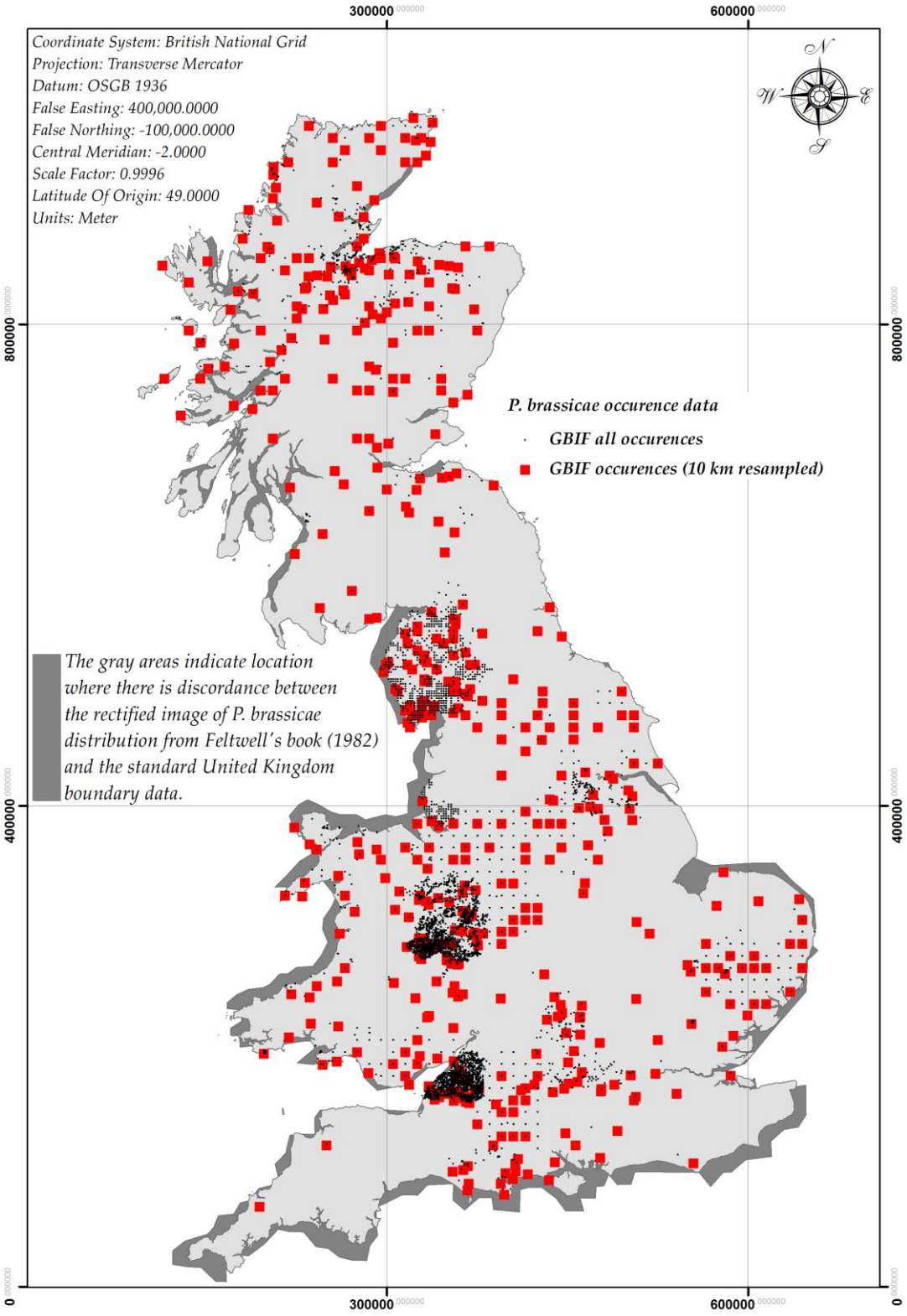
20th October, 1970 John HEATH

Data description given on the first page of the "Atlas of the Insects of The British Isles" (Heath, 1970) used to assess areas where P. brassicae *was absent within the time frame of the survey.*



Map of P. brassicae *distribution survey data extracted from the "Atlas of the Insects of The British Isles" (Heath, 1970).*

Appendix 7.2 Occurrence data accessed from the GBIF database



Appendix 7.3 Data sources

1) Description of data sources

P. brassicae occurrence data for New Zealand and the current surveillance methodology used for its eradication in New Zealand were accessed from agresearch (Phillips *et al.*, 2013).

P. brassicae occurrence data for United Kingdom from the GBIF[36] database were provided by various institutions that uploaded data to the GBIF portal and their list is given in the next page.

Additional occurrence data dating from early 1940's has been extracted from the atlas of insects of the British Isles prepared by Heath (1970) atlas, both the provisional atlas downloaded from the Open Research Archive of Natural Environment Research Council[37] website as well as the final version printed in Feltwell's (1982) book has been used to digitize *P. brassicae* occurrence points.

The standard United Kingdom boundary data[38] used to rectify the scanned *P. brassicae* atlas was downloaded from the United Kingdom Ordnance Survey website according to the license agreement given here http://www.ordnancesurvey.co.uk/docs/licences/os-opendata-licence.pdf.

The United Kingdom provincial data used to highlight administrative boundaries where unusual high resolution sampling of *P. brassicae* was undertaken were downloaded from Natural Earth[39] open source GIS data portal

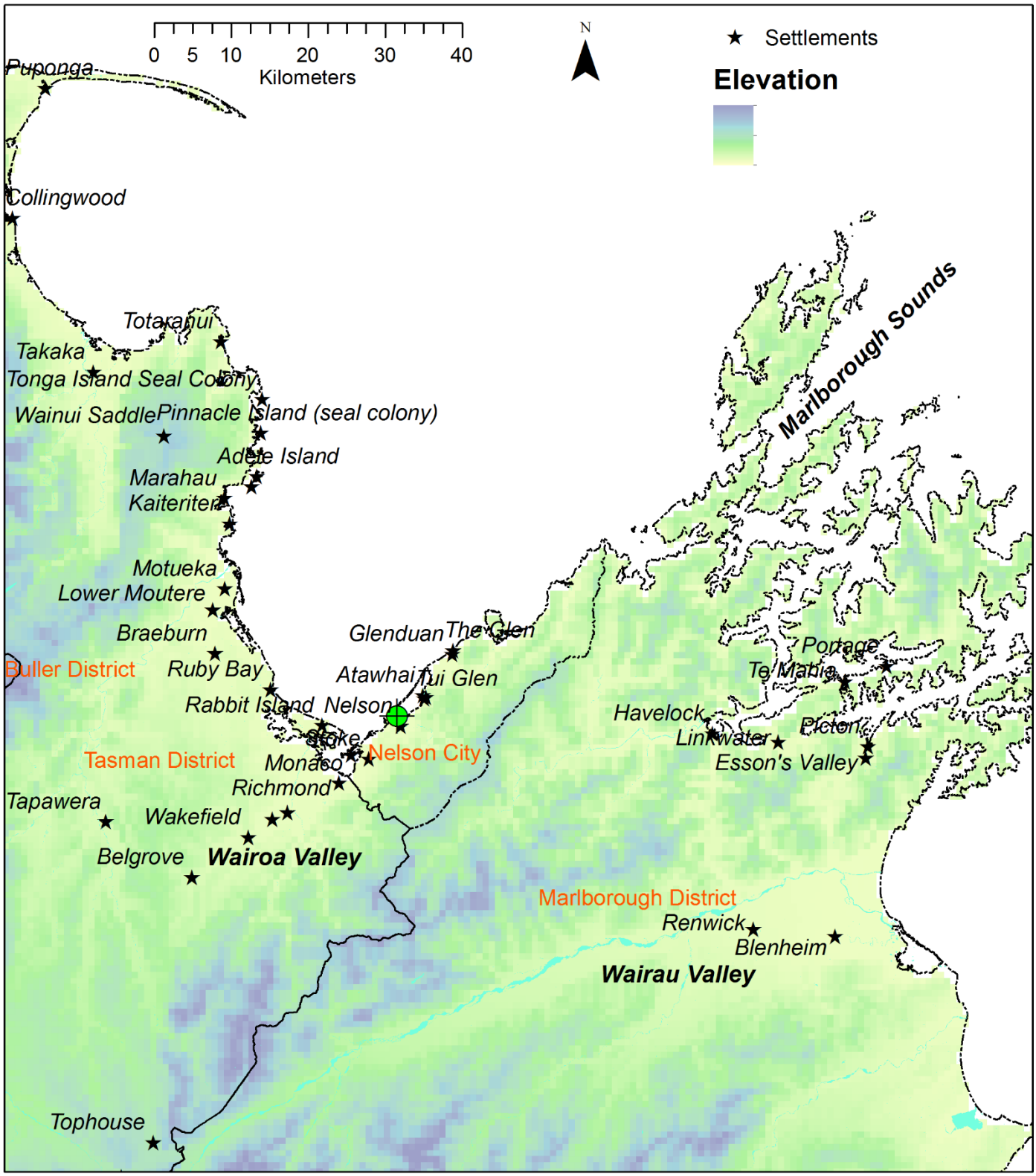
[36] http://www.gbif.org/
[37] NERC Open Access Research Archive (NORA) Available: http://nora.nerc.ac.uk/
[38] Contains UK Ordnance Survey data © Crown copyright and database right 2013
[39] Natural Earth© 2014, http://www.naturalearthdata.com/downloads/10m-cultural-vectors/

2) List of GBIF data sources

Data publisher	Dataset
UK National Biodiversity Network	Bristol Regional Environmental Records Centre - BRERC October 2009
UK National Biodiversity Network	BTCV Scotland - BTCV wildlife counts recording workshops
UK National Biodiversity Network	Cambridgeshire & Peterborough Environmental Records Centre - CPERC Recorders day at Waterbeach barracks and airfield
UK National Biodiversity Network	Countryside Council for Wales - Welsh Invertebrate Database (WID)
UK National Biodiversity Network	Countryside Council for Wales - Welsh Peatland Invertebrate Survey (WPIS)
UK National Biodiversity Network	Cumbria Biodiversity Data Centre - Cumbria Biodiversity Data Centre. Lepidoptera Observation Records. Pre-2010 for Cumbria
UK National Biodiversity Network	Dorset Environmental Records Centre - Dorset SSSI Species Records 1952 - 2004 (Natural England)
UK National Biodiversity Network	Hertfordshire Biological Records Centre - Hertfordshire Wildlife Site Monitoring Surveys (incomplete)
UK National Biodiversity Network	Highland Biological Recording Group - HBRG Insects Dataset
UK National Biodiversity Network	Humber Environmental Data Centre - Humber Environmental Data Centre - Non Sensitive Records from all taxonomic groups
UK National Biodiversity Network	Lothian Wildlife Information Centre - Lothian Wildlife Information Centre Secret Garden Survey
UK National Biodiversity Network	Merseyside BioBank - Merseyside BioBank Active Naturalists (verified)
UK National Biodiversity Network	Merseyside BioBank - North Merseyside Insects (verified)
UK National Biodiversity Network	National Trust - Anglesey Abbey wildlife species data held by The National Trust.
UK National Biodiversity Network	National Trust - Hatfield Forest species data held by The National Trust.
UK National Biodiversity Network	National Trust - Ickworth species data held by The National Trust.
UK National Biodiversity Network	National Trust - Sutton Hoo species data held by The National Trust.
UK National Biodiversity Network	National Trust - Wicken Fen nature reserve species data held by The National Trust
UK National Biodiversity Network	National Trust for Scotland - NTS Properties Species Records 1800-2013
UK National Biodiversity Network	Natural England - Invertebrate Site Register - England.
UK National Biodiversity Network	North & East Yorkshire Ecological Data Centre - North and East Yorkshire Ecological Data Centre - Non-sensitive Records from all taxonomic groups.
UK National Biodiversity Network	North Ayrshire Countryside Ranger Service - Species within North Ayrshire from 1984 - Present
UK National Biodiversity Network	Nottinghamshire Biological and Geological Records Centre - UK abstract from Nottingham City Museums & Galleries (NCMG) Insect Collection Baseline database
UK National Biodiversity Network	Open Mosaic Habitat Survey Group - Invertebrates recorded during Open Mosaic Habitat survey in England and Wales (2012)

Data publisher	Dataset
UK National Biodiversity Network	Outer Hebrides Biological Recording Project - OHBRP Insects Dataset - Outer Hebrides
UK National Biodiversity Network	Record the Biodiversity Information System for Cheshire Halton Warrington and the Wirral - RECORD Butterfly data up to current day
UK National Biodiversity Network	Rotherham Biological Records Centre - Rotherham Biological Records Centre - Non-sensitive Records from all taxonomic groups
UK National Biodiversity Network	Royal Horticultural Society - RHS monitoring of native and naturalised plants and animals at its gardens and surrounding areas
UK National Biodiversity Network	Scottish Borders Biological Records Centre - SWT Scottish Borders Local Wildlife Site Survey data 1996-2000 - species information
UK National Biodiversity Network	Seil Natural History Group - SNHG Biological Records Dataset
UK National Biodiversity Network	Sheffield Biological Records Centre - Sheffield Biological Records Centre- Non-sensitive Records from all taxonomic groups.
UK National Biodiversity Network	Shire Group of Internal Drainage Boards - Shire Group IDB species data 2004 to present
UK National Biodiversity Network	Shropshire Ecological Data Network - Shropshire Ecological Data Network Database
UK National Biodiversity Network	South East Wales Biodiversity Records Centre - CCW Regional Data : South East Wales Non-sensitive Species Records
UK National Biodiversity Network	Staffordshire Ecological Record - SER Site-based Surveys
UK National Biodiversity Network	Suffolk Biological Records Centre - Suffolk Biological Records Centre (SBRC) dataset
UK National Biodiversity Network	Thames Valley Environmental Records Centre - Local Wildlife Site Surveys Berkshire
UK National Biodiversity Network	Tullie House Museum - Tullie House Museum Natural History Collections.
UK National Biodiversity Network	Wiltshire and Swindon Biological Records Centre - Wiltshire & Swindon Site-based Survey Records
UK National Biodiversity Network	Yorkshire Wildlife Trust - Yorkshire Wildlife Trust - Non-sensitive records from all taxonomic groups

Appendix 7.4 Reference map of place names in the study area (Chapter 7)



Appendix 8.1 Research outputs

8.1.1 Conference abstracts and oral presentations

Senay, S. D. (2014). *Progress using SDM's for biosecurity decision making. P*resented at the Tripartite International Collaborative Research Initiative Knowledge Engineering and Discovery Research Institute (KEDRI) Auckland University of Technology, Shanghai Jao Tong & Xinjiang Universities China, Bio-Protection Research Centre, Lincoln, New Zealand.

Senay, S. D., Worner, S. P., & Ikeda, T. (2013). *Improved pseudo-absence selection technique for species distribution models.* Presented at the Bio-Protection Research Seminar Series, Lincoln University, Lincoln.

Senay, S. D., & Worner, S. P. (2013). *Why do models predict differently for the same species/location? (Awarded 2nd prize).* Presented at the Lincoln University Post-Graduate Conference, Lincoln University, Lincoln.

Senay, S. D., & Worner, S. P. (2013). *Correlative species distribution models-issues and solutions.* Presented at the B3 Pest risk modelling and mapping workshop for biosecurity. Plant & Food CRI, Lincoln.

Worner, S. P., **Senay, S. D.**, Khandan, H. A. N., & Lustig, A. (2013). *Characterizing the likelihood of establishment and spread on invasive pests on the post border pathway: current issues challenges and potential for hybrid or integrated models.* Paper presented at The Second International Congress on Biological Invasions, Qingdao, China.

Senay, S. D., Worner, S. P., & Ikeda, T. (2012). *A novel three-step pseudo-absence generation method with ecological, spatial and environmental aspects of species requirements considered.* Paper presented at the 8th International Conference on Ecological Informatics: Ecological Informatics for Biodiversity and conservation Biodiversity and Conservation., Brasilia, Brazil.

Senay, S. D., Worner, S. P., & Ikeda, T. (2012). *A novel three-step pseudo-absence selection method that balances environmental and geographical spaces.* Paper presented at the Pest Risk Modelling and Mapping workshop VI, Tromsoe, Norway.

Senay, S. D., Worner, S. P., & Ikeda, T. (2012). *Species distribution models and risk assessment.* Presented at the On Campus Event Relative risk of Augmented Pest Control, Bio-Protection Research Centre, Lincoln. The Australian and New Zealand organisation of the Society for Risk.

Senay, S. D., & Worner, S. P. (2012). *Spatial distribution models from the truth to the whole truth. (Invited)* Presented at the Canterbury statistics Open day, Canterbury University, Christchurch, New Zealand.

Senay, S. D. (2012). *Modelling alien invasive species-landscape interactions using high resolution spatially explicit models.* Presented at a Collaboration meeting between the Bio-Protection Research Centre and Plant and Food Research Centre, Te Puke, New Zealand.

Senay, S. D. (2011). *Species distribution models used in biosecurity. Insect invasion team research presentation.* Presented at the Tripartite International Collaborative Research Initiative Knowledge Engineering and Discovery Research Institute (KEDRI) Auckland University of Technology, Shanghai Jao Tong & Xinjiang Universities China, Bio-Protection Research Centre, Lincoln, New Zealand.

Senay, S. D. (2011). *Winning the war against alien bugs: new tools and tactics (Thr3sis competition).* Presented at the Thr3sis competition, Lincoln, New Zealand.

Senay, S. D., & Worner, S. P. (2011). *Using non-linear dimension reduction methods for multi-sourced, multi-format and multi-temporal geo-environmental predictors.* Paper presented at the Pest Risk Modelling and Mapping workshop V: Pest risk in a changing world, Fort Collins, USA.

Senay, S. D., & Worner, S. P. (2011). *Using non-linear dimension reduction methods for multi-sourced, multi-format and multi-temporal geo-environmental predictors.* Presented at the Lincoln University Post Graduate Conference, Lincoln, New Zealand.

Worner, S. P., Ikeda, T., Wang, D., **Senay, S. D.**, & Khandan, H. A. N. (2011). *Computational Intelligence and modelling in applied ecology.* Paper presented at the Computational Intelligence: methods systems and applications for ecological and environmental modelling in China and New Zealand, Auckland, New Zealand.

8.1.2 Publication and Research reports (Internal & external)

Journal publications and conference proceedings

Senay, S. D., Worner, S. P., & Ikeda, T. (2013). Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modelling. *PLoS ONE, 8*(8), e71218. doi:10.1371/journal.pone.0071218

Worner, S.P., Lankin, G., Lustig, A., Narouei Khandan, H.A., 1 **Senay, S.D.** (in Press) Being better than average: the application of computational intelligence in pest management and biosecurity. Pp xx-xx In RM Beresford, KJ Froud, JM Kean and SP Worner (Eds). Proceedings of New Zealand Plant Protection Society Symposium, The plant protection data toolbox: On beyond t, F and χ. New Zealand Plant Protection Society Inc., New Zealand

Internal & external reports

Logan, D., **Senay, S. D.**, & Khandan, H. A. N. (2013). *Habitat suitability predictions for selected glasshouse biological control agents using Maxent and Multi Modelling*. (SPTS No.8061)

Senay, S. D. (2010). *Modelling alien invasive species-landscape interactions using high resolution spatially explict models (15th month report)*. Lincoln, New Zealand.

Senay, S. D. (2010). *Modelling alien invasive species-landscape interactions using high resolution spatially explict models (Research proposal)*. Lincoln, New Zealand.

Appendix 8.2 Author contributions to the manuscripts associated with chapter 3 of this thesis

Results of Chapter 3 – published as:

> **Senay, S. D.,** Worner, S. P., & Ikeda, T. (2013). Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modelling. PLoS ONE, 8(8), e71218. doi:10.1371/journal.pone.0071218

Senay, S. D. – developed the methodology, designed and performed the research, collected data, analysed and interpreted the data, wrote the manuscript, prepared figures and table, submit the manuscript and addressed the revisions.

Worner, S.P. – advised on the use of the multi-model framework used in the research, provided research advices, provided assistance with manuscript preparation, helped to address the revisions, ensured the funding and provided editorial help.

Ikeda, T. - advised on the use of the multi-model framework used in the research, provided statistical advice.