Invasive Species Distribution Models: An Analysis of Scale, Sample Selection Bias, Transferability and Prediction

by

Jennifer Elisabeth Weaver

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Department of Geography and Program in Planning
University of Toronto
© Copyright by Jennifer Elisabeth Weaver 2012
Invasive Species Distribution Models:
An Analysis of Scale, Sample Selection Bias,
Transferability and Prediction

Jennifer Elisabeth Weaver
Doctor of Philosophy
Department of Geography and Program in Planning
University of Toronto
2012

Abstract

Species distribution models must balance the need for model generality with that for precision and accuracy. This is critical when modelling range-expanding species such as invasive species. Given the increased use of species distribution models to study invasive species-landscape relationships, a better understanding of the effect of spatial scales, sampling biases, model transferability and discrepancies between different models’ future predictions is necessary. This dissertation addresses these knowledge gaps using mute swans (*Cygnus olor*) as a case study species. I specifically examine mute swan’s distributions in parts of their native range of Britain and their non-native range of Ontario, Canada. I first investigate which environmental variables at which spatial scales best explain mute swan’s distribution in its non-native range. Second, I perform a sample selection bias study to examine predictive accuracy when species distribution models are built using varying ranges of environmental variables and applied to broader spatial extents. Third, I examine the potential for, and limitations of model transferability between native and non-native regions. Finally, I use two different modelling approaches and three
different climate change and land use change scenarios to predict future mute swan habitat suitability. The results indicate that (1) models with better predictive accuracy include environmental variables from multiple ecologically-meaningful scales and measured at spatial extents that include a broad range of environmental variable values; (2) models can exhibit asymmetrical transferability; (3) climate change will facilitate mute swan range expansion in the future more than land use change; and (4) mute swans are often found near urban waterbodies. When modelling invasive species distributions, I suggest that ecologists consider: (I) spatial scale of the underlying landscape processes and species’ use of the landscape; (II) variability and range of each environmental variable used for building models; and (III) stage of establishment of the invasive species.
Acknowledgments

I would like to sincerely thank my supervisors, Dr. Tenley Conway and Dr. Marie-Josée Fortin for their invaluable advice and support throughout my PhD. In addition, I would like to thank the members of my committee: Dr. Sharon Cowling, Dr. Don Boyes, Dr. Sarah Finkelstein and Dr. Andy Kenney. Peter Lack at the British Trust for Ornithology and Denis Lepage at Bird Studies Canada were also invaluable for supplying data and answering data-related questions. I would also like to thank the official sponsors of the Ontario Breeding Bird Atlas (Bird Studies Canada, Canadian Wildlife Service, Federation of Ontario Naturalists, Ontario Field Ornithologists, and Ontario Ministry of Natural Resources) for supplying Atlas data, and to the thousands of volunteer participants who gathered data for the project. I would also like to acknowledge the Natural Science and Engineering Research Council of Canada, Ontario Graduate Scholarship in Science and Technology, the Department of Geography and the University of Toronto, among others for funding my research, conference attendance and other academic endeavours.

My family, friends and colleagues have been phenomenal in their support of me while writing this dissertation. I would especially like to acknowledge the support by my family, my friends, and my colleagues in the House lab, Landscape Ecology lab and Cowling lab.
# Table of Contents

Acknowledgments .......................................................................................................................... iv

List of Tables ................................................................................................................................... x

List of Figures ............................................................................................................................... xii

Chapter 1   Introduction ...................................................................................................................1

1.1 Invasive Species...................................................................................................................1

1.2 Case Study Species – Mute Swans ......................................................................................1

1.3 Study Site: Maps of Water and Urban Land Cover .............................................................3

1.4 Species Distribution Models and Glossary of Terms ..........................................................6

1.5 Knowledge Gaps..................................................................................................................8

  1.5.1 Scale of Study Data..................................................................................................9

  1.5.2 Issues of Range of Environmental Variables and Transferability in Species
       Distribution Models .................................................................................................9

  1.5.3 Climate change, land use change and variability in predictions between
       species distribution models ....................................................................................11

  1.5.4 Invasive Mute Swan Ecology ................................................................................12

1.6 Research Objectives...........................................................................................................12

1.7 Dissertation Structure and Publication Information ..........................................................13

Chapter 2  An invasive species’ relationship with environmental variables changes when
measured at different spatial scales ..............................................................................................16

  2.1. Introduction................................................................................................................16

  2.2. Methods.....................................................................................................................18

    2.2.1. Study Site ...........................................................................................................18

    2.2.2. Case Study Species: Mute Swans ......................................................................19

    2.2.3. Relevant Ecological Scales ................................................................................21
3.3.1. Comparison of models ...........................................................................................47
3.3.2. Predictive accuracy of models ...............................................................................47
3.3.3. Range of environmental variables .........................................................................48
3.3.4. Variable importance ...............................................................................................52
3.4. Discussion ..........................................................................................................................52
3.4.1. Models – all variables vs. climate vs. land cover .................................................53
3.4.2. Predictive accuracy ................................................................................................53
3.4.3. Environmental variables ........................................................................................54
3.5. Conclusions........................................................................................................................55

Chapter 4  Assessing Model Transferability: Modelling Recently Established Invasive Species Distributions in Native and Non-Native Ranges ............................................................57
4.1. Introduction...................................................................................................................57
4.2. Methods.........................................................................................................................59
4.2.1. Study Sites .............................................................................................................59
4.2.2. Mute Swan Distribution Datasets ..........................................................................60
4.2.3. Predictor Variables.................................................................................................61
4.2.4. Models....................................................................................................................62
4.2.5. Statistical Analysis.................................................................................................63
4.3. Results............................................................................................................................65
4.4. Discussion ....................................................................................................................71
4.5. Conclusions....................................................................................................................75

Chapter 5  Impacts of Climate Change and Land Use Change on an Invasive Species’ Distribution ............................................................................................................................77
5.1. Introduction...................................................................................................................77
5.2. Methods.........................................................................................................................80
5.2.1. Study Site...............................................................................................................80
Chapter 5: Models of Mute Swan Distribution

5.2. Mute Swan Distribution Data ..........................................................................................80
- Mute Swan Distribution Data......................................................................................................80
- Sampling Effort.........................................................................................................................81
- Models: MaxEnt and Generalized Linear Models .....................................................................81
- Predictor Variables.....................................................................................................................82
- Climate Change Scenarios .......................................................................................................83
- Urbanization Scenarios .............................................................................................................84
- Habitat Suitability Prediction Mapping ..................................................................................85

5.3. Results ................................................................................................................................85

5.4. Discussion ..........................................................................................................................100
- Response to Climate Change ..................................................................................................100
- Response to Urbanization .......................................................................................................101
- Uncertainty in Predictions ......................................................................................................102
- Biotic Considerations .............................................................................................................102

5.5. Conclusions ......................................................................................................................103

Chapter 6: Summary and Synthesis ..........................................................................................104

6.1. Summary ..........................................................................................................................104
- The scale at which environmental variables best explain mute swan distribution ..................105
- The impact of sample selection bias and spatial extent on model predictive accuracy ..............106
- Model transferability between native and non-native ranges ..................................................106
- Using species distribution models to predict future distributions under climate change and land use change scenarios .................................................................108

6.2. Synthesis ..........................................................................................................................109
- Consideration of scale in species distribution models ..............................................................109
- Consideration of range of environmental variables in species distribution models ..................109
6.2.3. Consideration of transferability of species distribution models .........................110

6.2.4. Consideration of predictions of species distributions when accounting for climate change and land use change ..................................................................................110

6.2.5. Contributions to mute swan ecology..................................................................111

6.3. Future Research Directions.......................................................................................111

6.3.1. Species distribution models and connectivity....................................................112

6.3.2. Dispersal Rate .....................................................................................................112

6.3.3. Multi-Species and Larger Spatial Extent Transferability Studies.....................113

References.........................................................................................................................114
List of Tables

Table 1-1 Glossary of terms used in this dissertation……………………………………………… 7

Table 2-1 Categories of environmental variables potentially affecting mute swan’s distribution ……………………………………………………………………………………………………... 23

Table 2-2 Eight single-scale and three composite-scale models explaining mute swan’s distribution in its non-native range, and the environmental variables included in each…………………………………………………………………………………………………….………… 26

Table 2-3 Akaike weights of all models at all scales of analysis…………………………………… 29

Table 2-4 Akaike weights for all single-scale models……………………………………………… 29

Table 2-5 Area under the receiver operator curve values for all models with Akaike weights > 0.00…………………………………………………………………………………………..... 30

Table 2-6 Kappa values for all models with Akaike weights > 0.00…………………………….. 30

Table 3-1 Environmental variables included in theoretical models to explain mute swan distribution in native range of Britain (with transformations in parentheses)………………... 40

Table 3-2 Akaike weights (evaluated within sub-extents) for each model built at seven spatial sub-extents…………………………………………………………………………………………….. 47

Table 3-3 AUC values of the best models of the best run per sub-extent and AUC, Kappa and sensitivity values based on application of each model to the entire study extent……………… 48

Table 3-4 Percent loss of explained deviance when individual variables are removed from the model of each sub-extent……………………………………………………………… 52
Table 4-1 Theoretical models explaining mute swan (*Cygnus olor*) distribution in its native (Britain) and non-native (Ontario) ranges (with transformations of environmental variables in parentheses).............................................................................................................................................................. 63

Table 4-2 Mean AUC values of models validated internally................................................................. 65

Table 4-3 Evaluation of Land cover + Climate, Habitat + Climate and Land cover final models: Akaike weights............................................................................................................................................................................................................... 66

Table 4-4 Evaluation of Land cover + Climate + Elevation, Habitat + Climate and Land cover final models: Akaike weights............................................................................................................................................................................................................... 66

Table 4-5 Evaluation of model transferability between native and non-native ranges (comparing internal and external model validation measures such as AUC and maximum Kappa).............. 67

Table 4-6 The percent loss of explained deviance when individual environmental variables are removed from the Land Cover + Climate + Elevation models of Britain (Nagelkerke’s $R^2$ value = 0.350) and Ontario (Nagelkerke’s $R^2$ value = 0.508)........................................................................................................................................................................ 71

Table 5-1 Permutation importance of each environmental variable to the built model as determined by *MaxEnt*........................................................................................................................................................................ 86

Table 5-2 Percent loss of explained deviance in the built model when individual environmental variables are removed from the full GLM with a Nagelkerke’s $R^2$ value of 0.453............... 88
List of Figures

**Figure 1-1** Percent water cover (a) and urban land cover (b) distribution per 10 km grid cell at study site of Britain................................................................. 4

**Figure 1-2a** Percent water cover distribution per 10 km grid cell at study site of southern Ontario........................................................................................................... 5

**Figure 1-2b** Percent urban land cover distribution per 10 km grid cell at study site of southern Ontario........................................................................................................... 5

**Figure 2-1** Map of study site of southern Ontario (Canada) illustrating urban land cover and mute swan distribution (presence and pseudo-absence data) from 2001-2005 (Cadman *et al.*, 2007; Ontario Ministry of Natural Resources, 2008)........................................................................................................... 19

**Figure 2-2** Three main ecological scales were used to measure landscape and climate variables to include in the species distribution model; 140m, 3000m and 8000m............................................. 22

**Figure 2-3** Coefficients and corresponding standard errors of the Global single-scale model applied at the 140m, 3000m and 8000m ecological zones................................................................. 30

**Figure 3-1** Study site of Britain illustrating grid cells that were not visited in the survey, absence and presence grid cells.................................................................................. 38

**Figure 3-2** Sub-extents used for geographically dividing full extent (a) of study site. These were divided using freshwater regions – Freshwater 402 and Freshwater 404 (b), and mute swan presence density – Outer Core and Inner Core (c)........................................................................................................... 42

**Figure 3-3** Flow chart of data partitioning and dataset creation for each sub-extent................. 44

**Figure 3-4** Mean and range of the values of each environmental variable used per sub-extent.. 49
Figure 3-5 Histogram of percent urban land cover variable for all four sub-extents and full extent……………………………………………………………………………………………………50

Figure 3-6 Histogram of precipitation variable for all four sub-extents and full extent………50

Figure 3-7 Histogram of mean temperature of warmest quarter variable for all four sub-extents and full extent…………………………………………………………………………………………………………………………………………..51

Figure 3-8 Histogram of mean temperature of coolest quarter variable for all four sub-extents and full extent…………………………………………………………………………………………………………………………………………..51

Figure 4-1 Predicted probability of mute swan presence in its native range of Britain using: (a) the LCC and Habitat + Climate model built in mute swan’s native range of Britain and (b) the LCC and Habitat + Climate model built in mute swan’s non-native range of Ontario. Differences in predicted probabilities determined by subtracting Ontario’s probability values from Britain’s probability values illustrated in (c)…………………………………………………………………………………………………………………………………………..68

Figure 4-2 Predicted probability of mute swan presence in its non-native range of Ontario using: (a) the LCC and Habitat + Climate models built in mute swan’s non-native range and (b) the LCC and Habitat + Climate model built in mute swan’s native range. Differences in predicted probabilities determined by subtracting Britain’s probability values from Ontario’s probability values illustrated in (c)…………………………………………………………………………………………………………………………………………..70

Figure 5-1 Importance of environmental variables for current mute swan distribution in Ontario as determined by a jackknife of the regularized training gain in MaxEnt…………………………86

Figure 5-2 Variable response curves when each environmental variable is used individually in MaxEnt including (a) mean temperature of coolest quarter (bio11), (b) annual precipitation (bio12), (c) percent urban land cover (perurb), and (d) percent water cover (perwat)…………87
Figure 5-3 Maps of predicted suitable habitat for mute swans at present (2001-2005) using (a) MaxEnt with a regularization factor of 1, (b) MaxEnt with a regularization factor of 4, and (c) a GLM. Current mute swan presence locations used to train the models are also shown. 90

Figure 5-4 Maps of predicted probabilities of suitable habitat for mute swans in 2050 using the A2 climate change scenario using (a) MaxEnt with a regularization factor of 1, (b) MaxEnt with a regularization factor of 4, and (c) a GLM. 92

Figure 5-5 Maps of predicted probabilities of suitable habitat for mute swans in 2050 using the urbanization scenario using (a) MaxEnt with a regularization factor of 1, (b) MaxEnt with a regularization factor of 4, and (c) a GLM. 94

Figure 5-6 Maps of predicted probabilities of suitable habitat for mute swans in 2050 using the A2 climate change and urbanization scenarios using (a) MaxEnt with a regularization factor of 1, (b) MaxEnt with a regularization factor of 4, and (c) a GLM. 96

Figure 5-7 Maps illustrating the limiting environmental variables for the (a) climate scenarios and (b) urbanization scenario. 98

Figure 5-8 Maps illustrating the levels of uncertainty in the predictions using the (a) climate scenario and (b) urbanization scenario. 99
Chapter 1
Introduction

1.1 Invasive Species

Invasive species tolerate a broad range of biotic and abiotic conditions, have few competitors, and can exhibit greater and more rapid population growth in novel ranges than they do in their native range (Martin, 1999; Sanford et al., 2003; Jakobs et al., 2004). The distribution and spread of invasive species is therefore of significant concern as they can become notably widespread and pervasive (Orians, 1986; Sax et al., 2007). Establishment of an invasive species can place stress on an ecosystem by decreasing ecosystem productivity, interfering with natural ecological processes by increasing competition for resources, and by displacing native species (Martin, 1999; Davies and Shely, 2007). Indeed, exotic invasive species have a unique relationship with the landscape that differs from that of native species as they often thrive in environments that are highly disturbed, such as the urban areas that native species frequently avoid (Orians, 1986). In many cases, little is known about invasive species’ relationships with the environment and dispersal patterns in non-native ranges (With, 2002). Yet to effectively plan and manage for future invasions and range expansion of invasive species, ecosystem managers must understand the distribution patterns and species-landscape-climate relationships of invasive species.

1.2 Case Study Species – Mute Swans

Mute swans are a practical case study species for studying issues of invasive species distribution modelling as there are thorough datasets that have been collected for mute swans both in their native and non-native ranges as part of the 1990 Mute Swan Survey and 2001-2005 Ontario Breeding Bird Atlas, respectively. Mute swans are easy to atlas as they are conspicuously large,

---

1 The definition of invasive species that I will use in my dissertation is that used by the IUCN Guidelines for the Prevention of Biodiversity Loss caused by Alien Invasive Species. They define an exotic invasive species as “an alien species which becomes established in natural or semi-natural ecosystems or habitat, is an agent of change, and threatens native biological diversity” (International Union for the Conservation of Nature, 2000).
contrastingly white and because the male mute swan often parades near the water’s edge in breeding season (Cadman et al., 1987). Mute swans are non-migratory birds endemic to Europe and Asia (Allin et al., 1987). They were first introduced to North America via the urban parks of New York City in the late 19th century, and were introduced to Ontario in 1958 (Petrie and Francis, 2003). Since then, they have escaped or have been released into the wild and continue to reproduce independently of human influence. Overall, little research has examined the relationship between mute swans and the landscape in its non-native range. Much of the previous work examining mute swan populations has focused on endocrinology, avian flu, eating habits, breeding, diseases, songs, toxicity, and lead poisoning (Nummi and Saari, 2003; Bailey et al., 2008; Hars, 2008).

The majority of studies that have considered mute swan habitat preferences have been conducted in mute swan’s native range of Europe and Asia (Wieloch et al., 2004), and of these, the majority have been performed in Britain or Western Europe (Brown and Brown, 1993; Kirby et al., 1994; Holm, 2002; Fuller et al., 2005; Mason et al., 2006). The mute swan population of the western Palearctic is estimated to be approximately 550 000 birds (Wieloch et al., 2004) and most studies of this population have concentrated on physiology and population dynamics. However, those studies that have focused on habitat preferences have revealed that there has been a noticeable change in mute swan’s relationship with the landscape in its native range as mute swans are now found more often in urbanized areas (Wieloch et al., 2004). For example, mute swans are more often found in urban stretches than rural stretches of rivers (Mason et al., 2006), exhibit a higher tolerance to pollution (Holm, 2002) and are found in smaller bodies of water characteristic of urban areas (Wieloch et al., 2004). This may be because there is a growing dependence of the mute swan on food and habitat provided by humans and because they can tolerate the noise and pollution levels often found in built environs (Wieloch et al., 2004).

Mute swans are charismatic megafauna which have received growing attention in the Ontario media for their aggressive protection of their nests (Conover and Kania, 1994; Petrie and Francis, 2003). In Ontario, mute swans are currently found on the coasts of all of the Great Lakes with the exception of Lake Superior (Cadman et al., 2007). The Ontario population of mute swans has been given the distinction of being an invasive species due to its rapid population growth and
competition for food and habitat with native waterfowl (Petrie and Francis, 2003). The average mute swan directly consumes approximately 3.6 kilograms (wet weight) of submerged aquatic vegetation (SAV) per day and disturbs additional SAV during foraging, which reduces the resources available to other species (Kirby et al., 1994; Petrie, 2002). In addition to competition for resources, mute swans can act aggressively towards native species (Conover and Kania, 1994; Petrie and Francis, 2003). Black ducks, black skimmers, terns and tundra swans have all been negatively affected by mute swans either through competition for food or territory (Ciaranca et al., 1997; Petrie, 2002). Mute swans have often been associated with urban areas in their invasive range, and therefore will function as an excellent case study in which to examine the relationship between urban development and invasive species distribution and range expansion.

In the past decade, the mute swan population has grown rapidly in southern Ontario. The Ontario Breeding Bird Atlas estimated the 2005 population to be 2737 individuals, which was almost twice as high as the estimated population in 2002 (Cadman et al., 2007). Petrie and Francis (2003) estimate that the Great Lakes mute swan population could be increasing at a rate of 10 to 15% per year and grow to be 16 000 in the next 40 to 50 years based on mute swan territory size and available wetland habitat (Petrie, 2002; Petrie and Francis, 2003). This would be a notable increase from their population size in 2005 (Cadman et al., 2007). Though considered to be invasive, mute swans are still protected under the Canadian Migratory Bird Conservation Act through a loophole in which all species of swans are protected (Petrie, 2002). In Ontario, the range of the mute swan and its native congener, the trumpeter swan, rarely overlap (Cadman et al., 2005) and it is currently unknown as to whether these populations will compete in the future. As mute swans do not yet occupy all possible territories in Ontario, the predictions made by this dissertation hold the potential to be very useful for wildlife managers in terms of preventative management.

1.3 Study Site: Maps of Water and Urban Land Cover

The following maps illustrate the two main study areas for this dissertation, Britain (Figure 1-1a and 1-1b) and Ontario (Figure 1-2a and 1-2b). For each study area, the main land cover types
that are particularly important for this dissertation, percent water and percent urban land cover are illustrated.

![Figure 1-1](image)

**Figure 1-1** Percent water cover (a) and urban land cover (b) distribution per 10 km grid cell at study site of Britain.
Figure 1-2a Percent water cover distribution per 10 km grid cell at study site of southern Ontario.

Figure 1-2b Percent urban land cover distribution per 10 km grid cell at study site of southern Ontario.
1.4 Species Distribution Models and Glossary of Terms

Species distribution models assess the correlative relationships between the environment and species’ locations within that environment (Franklin, 2009). In addition to many different types of species distribution models, there are many different methods of assessing goodness-of-fit and predictive accuracy of species distribution models; those that are used in this dissertation are defined here (Table 1-1). As there are a growing number of studies using species distribution models to model invasive species, it is prudent to further examine different facets of invasive species distribution modelling. Modelling of an invasive species is a specific type of problem in terms of predictions as it involves range-expanding species where the landscape-climate-relationship found in the native range may differ from that found in the non-native range (Ficetola et al., 2007; Beaumont et al., 2009). Furthermore, modelling a species that: (1) is not at equilibrium with the environment; (2) may have different species-landscape-climate relationships between the native and non-native ranges; and (3) can adapt to or establish in marginal habitats, becomes more complicated (Elith et al., 2010; Václavík and Meentemeyer, 2012). In terms of the success of applying species distribution models to invasive species, there remain many unanswered questions (Welk, 2004; Mau-Crimmins et al., 2006; Broennimann et al., 2007; Beaumont et al., 2009; Elith et al., 2010), to which this dissertation will contribute to answering.
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
<th>Source(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akaike information criterion (AIC)</td>
<td>AIC is an evaluation of the fit of a model, and effectively penalizes for the number of parameters used, with lower AIC values indicating better fit.</td>
<td>Rushton <em>et al.</em>, 2004</td>
</tr>
<tr>
<td>Akaike information criterion c (AICc)</td>
<td>AICc is a variation of AIC which was developed for datasets in which the ratio of sample size to the number of parameters is less than 40. The lowest AICc values indicate the best model fit.</td>
<td>Burnham and Anderson, 1998</td>
</tr>
<tr>
<td>Akaike weight</td>
<td>Akaike weights are calculated solely based on AICc values and are used to determine the weight of each model for model selection. Models with higher weights have the best fit.</td>
<td>Burnham and Anderson, 1998</td>
</tr>
<tr>
<td>Area under the curve (AUC) of the receiver-operating characteristic plot</td>
<td>AUC values are predictive accuracy measures independent of threshold values. Models with AUC values greater than 0.9 exhibit excellent predictive accuracy, while values of 0.8-0.89 and 0.7-0.79 indicate good and fair predictive accuracy respectively.</td>
<td>Swets, 1988; Manel <em>et al.</em>, 2001</td>
</tr>
<tr>
<td>Generalized linear model</td>
<td>Generalized linear models are a form of regression in which the dependent variable is linked to independent variables via a link function (in this case logit). They require both presence and absence data.</td>
<td>Lutolf <em>et al.</em>, 2006</td>
</tr>
<tr>
<td>Maximum Kappa</td>
<td>Kappa values are predictive accuracy measures that are dependent on a threshold. Maximum Kappa values are determined by finding the highest Kappa value for all thresholds between 0 and 1.</td>
<td>Randin <em>et al.</em>, 2006</td>
</tr>
<tr>
<td><strong>MaxEnt</strong></td>
<td><strong>MaxEnt</strong> is a machine-learning model based on a maximum entropy framework where, instead of using a regressive relationship, the model looks for a solution which is the most uniform between the species and the environmental variables.</td>
<td>Phillips et al., 2006</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Minimized difference between sensitivity and specificity (MDSS) threshold</td>
<td>Presence thresholds are calculated based on the minimized difference between sensitivity and specificity values.</td>
<td>Liu et al., 2005</td>
</tr>
<tr>
<td><strong>Nagelkerke’s $R^2$ ($D^2$)</strong></td>
<td>A pseudo $R^2$ estimate for logistic models used to evaluate how much of the variance each model explains. As with $R^2$ values, the possible range of values is between 0 and 1.</td>
<td>Graf et al., 2006</td>
</tr>
<tr>
<td><strong>Sensitivity</strong></td>
<td>Sensitivity values are concerned with the accurate prediction of presence points and reflect the proportion of presence points accurately predicted as presences.</td>
<td>Liu et al., 2005</td>
</tr>
<tr>
<td><strong>Specificity</strong></td>
<td>Specificity values are concerned with the accurate prediction of absence points and reflect the proportion of absence points accurately predicted as absences.</td>
<td>Liu et al., 2005</td>
</tr>
</tbody>
</table>

### 1.5 Knowledge Gaps

This dissertation addresses knowledge gaps that exist in modelling invasive species distributions, including (1) invasive species-landscape relationships; (2) implications of varying the scale at which environmental variables are measured; (3) influence of bias in range of environmental variables used to build models; (4) potential for transferability of models between regions (specifically native and non-native ranges) and (5) predictions of future distributions using different modelling methods and scenarios of both climate change and land use change.
1.5.1 Scale of Study Data

A concern with studying any species’ distribution is the scale at which the environmental variables are measured (Meyer and Thuiller, 2006; Elith and Leathwick, 2009). While land cover, climate variables and other environmental variables may be available at varying scales, the researcher must independently determine at what scale these variables should be measured. In the case of presence-absence grid cell data, this question is often answered by default, by using the resolution of the grid cell. However, in the case of point data, the landscape could be measured at any scale from the smallest resolution of any of the variables to the entire study extent. Thus far, there is not one ecological scale of measurement rule that can be applied across species (Elith and Leathwick, 2009). While there have been some local scale studies performed on mute swans, there have been few that have focused on the regional and landscape scales (Kirby et al., 1994). A landscape scale study allows for a better examination of the range of mute swans and its correlated landscape characteristics at a larger extent, as the processes and interactions measured at larger scales can differ from those observed at fine scales. In this dissertation, I address the following questions: (1) Which single-scale model best explains mute swan distribution and at what scale? (2) Are composite models comprised of environmental variables measured at different scales better at explaining mute swan distribution? and (3) What is the relationship between environmental variables and the invasive mute swan’s current distribution and how is this relationship affected by the spatial scale of analysis? This will not only offer further insight into mute swan ecology in its non-native range, but it will also address the question of the role of scale in the modelling of any species’ distribution.

1.5.2 Issues of Range of Environmental Variables and Transferability in Species Distribution Models

Species distribution models are often built using data from one particular area and then used to predict distributions in other locations (Graf et al., 2006; Barbosa et al., 2009; Zanini et al., 2009). These other locations may be another part of the native range (Graf et al., 2006; Barbosa et al., 2009; Zanini et al., 2009), or in the case of exotic invasive species, could include newly invaded territory (Fitzpatrick et al., 2006; Mau-Crimmins et al., 2006). Often, the choice of data used to construct models is based on availability or feasibility for collection. For example, if data
are collected only where a species is highly concentrated, the built model is unlikely to capture marginal habitats and variability in species’ distributions. Thus, models built from this data will likely under-predict distributions. In this dissertation, I explore the relative success of building models using data collected from spatial extents that account for little variability, include marginal habitats or are limited to ideal habitats with higher densities of the studied species. I ask the following questions: (1) What are the effects of using limited sampling extents (and therefore limited ranges of environmental variables) to model species’ distributions? and (2) Are there general characteristics of sampled landscape that result in more generalizable models?

Furthermore, the relationship between a species and its landscape is also thought to differ between its native and non-native ranges due to differing competitors, resource availability and climates (Martin, 1999; Sanford et al., 2003; Jakobs et al., 2004). However, due to limited data in the non-native range, data from the native range may be relied upon to predict distributions and potential relationships in the non-native range. A growing number of studies compare a species’ relationship with the native landscape to that species’ relationship with its non-native range (Fitzpatrick et al., 2006; Broennimann et al., 2007; Hill et al., 2012; Wenger and Olden, 2012). In this light, recent studies have addressed the feasibility of model transferability between native and non-native ranges (Fitzpatrick et al., 2006; Broennimann et al., 2007; Wenger and Olden, 2012).

Transferability of a model can be deemed to be successful when its predictive accuracy of the distribution in the range where the model was not built is comparable to the predictive accuracy when validated in the range in which the model was built (Randin et al., 2006). Furthermore, the concept of successful transferability posits that the species-landscape-climate relationship does not significantly change between ranges, and that important variables that exert the most influence on the species’ distribution remain the same. This raises the following question: Can a species distribution model for the invasive mute swan be successfully transferred between parts of its native and non-native ranges?
1.5.3 Climate change, land use change and variability in predictions between species distribution models

As with all predictions of future conditions, there is a significant amount of uncertainty in predicting not only how the climate and land cover will change in the future, but how species will respond to these changes (Iverson et al., 2004; Araújo et al., 2005; Franklin, 2009). Species distribution models can be utilized to predict future distributions using the assumption that the current species-landscape-climate statistical relationships will remain constant (Wisz et al., 2008; Wiens et al., 2009; Elith et al., 2010). Many have predicted the effects of climate change and land use change separately on range extent of a species (Wiens et al., 2009; Loiselle et al., 2010; Adams-Hosking et al., 2011), but few have examined the effects of both, and furthermore compared the magnitude of each (Wisz et al., 2008b). While land use change can be disruptive to native species due to fragmentation and loss of habitat (Evans et al., 2009), land use change has been shown to benefit invasive species (Blair, 1996; Bartuszevige et al., 2006). That is, urban sprawl may actually be conducive to the range expansion of some invasive species that can thrive in highly disturbed landscapes (Smith, 2007). Climate change can result in range expansion, range shifts or range contraction depending on a species’ relationship with temperature and precipitation. I ask the following questions: (1) Does the invasive mute swan exhibit range expansion or range shift with climate change and/or land use change? and (2) Do different modelling approaches result in different predictions of future suitable habitats?

While many studies have predicted the effect of climate change on the ranges of native species (Pearson and Dawson, 2003; Bowman et al., 2005), few have examined its relationship with invasive species (Jarnevich and Stohlgren, 2009). Based on breeding birds in Ontario, a recent study concluded that avian breeding ranges in Ontario have not (on average) shifted northwards with a changing climate, however some species which are at the northern edge of their breeding range in Ontario have shown a range shift northwards (Cadman et al., 2007). In addition, Cadman et al. (2007) hypothesize that when examining species-landscape relationships at a smaller scale, land use change may be more important than climate change.
1.5.4 Invasive Mute Swan Ecology

While many studies have examined the correlation between avian richness or abundance and landscape structure (Earnst, 1994; Drolet et al., 1999; Gammon and Maurer, 2002; Pearce et al., 2007), including urban landscapes (Blair, 1996; Chace and Walsh, 2006; Donnelly and Marzluff, 2006; Clergeau and Quenot, 2007; Tratalos et al. 2007; Burger and Gochfeld, 2009; Evans et al., 2009; Melles et al., 2010), the majority of these studies have focused on native species. It has been previously established that invasive species often exhibit a unique relationship with the environment as they are often able to capitalize on marginal, highly-disturbed habitats such as urban areas, which are actually likely to meet invasive species’ resource needs (Orians, 1986). Mute swans have been extensively studied in their native range, especially in Britain, but few studies have been conducted that examine the habitat preferences and landscape relationships of mute swans in their non-native range (for exceptions, see Ciaranca et al., 1997; Sousa et al., 2008). A unique quality of mute swans is that they have been associated with urban areas in both their native and non-native ranges (Ciaranca et al., 1997; Wieloch et al., 2004) but no studies have been performed to compare relationships in these ranges. As mute swans are an invasive species in North America, and can have significant impacts on habitat and resources, conservation managers require further information about mute swan ecology in their non-native range, or at a minimum, confirmation that mute swan’s habitat preferences and species-landscape-climate relationships do not differ from that of their native range. To address these knowledge gaps, I examine which environmental variables determine mute swan distribution in both its native and non-native range.

1.6 Research Objectives

This dissertation examines the effects of varying certain parameters of species distribution models when applied to an invasive species. This dissertation’s purpose is to fill the aforementioned knowledge gaps that exist in species distribution modelling in relation to their application to invasive species in particular.
The major objectives of this dissertation are the following:

1. Determine whether single scale or composite scale models can best represent mute swan’s distribution;
2. Examine the effect of varying or limiting the range of environmental variables when building models for a native species;
3. Test the transferability of models between native and non-native ranges;
4. Compare predictions made by two different types of species distribution models;
5. Predict the relative impact of climate change and urbanization on an invasive species’ range;
6. Contribute to mute swan ecology knowledge and management efforts by specifically examining their relationship with the landscape and predicting areas of suitable habitat both at present and in the future.

1.7 Dissertation Structure and Publication Information

Chapter 1
Chapter 1 introduces the context to my research, discusses the current knowledge gaps that exist in this field that my dissertation will fill and presents my major research questions and objectives.

Chapter 2
An invasive species’ relationship with environmental variables changes when measured at different spatial scales
Chapter 2 assesses the effects of varying the scale at which environmental variables are measured when building species distribution models for point data. The spatial scales at which birds perceive the landscape are not fully understood, and yet we know that individuals select their habitat based on a range of life stage needs that are likely scale-specific. This chapter examines the effects of scale by measuring the landscape at three different scales, building single-scale and composite-scale models and determining which model (and at which scale) has the best fit and predictive accuracy in determining mute swan’s distribution. It gives insight into
the relative contributions of environmental variables that determine mute swan distribution in its non-native range, as well as the scale at which the mute swan may perceive the landscape in terms of establishment. The manuscript for this chapter is co-authored by Tenley M. Conway and Marie-Josée Fortin and is published in *Landscape Ecology*.

Chapter 3

**The effect of sample selection bias on species distribution models' predictions**

Chapter 3 examines the effects of changing the range of environmental variables by modifying spatial extent on the built model’s predictive accuracy of the entire study extent. Selection bias in the data used for models is prevalent but often not acknowledged nor examined for its effects on the built model. In this chapter, I examine the implications of varying extents and importance of including marginal habitat in building species distribution models. The manuscript for this chapter is co-authored by Marie-Josée Fortin and Tenley M. Conway and will be submitted to *Global Ecology and Biogeography*.

Chapter 4

**Assessing model transferability: Modelling recently established invasive species distributions in native and non-native ranges**

Chapter 4 examines the potential for model transferability between the native and non-native ranges of an invasive species. This chapter examines whether and how the relationship between a species and its native landscape and climate differs from its relationship with its non-native landscape and climate. I examine the accuracy and predictive capacity of transferring models between ranges and the implications of this for species distribution predictions in non-native ranges. This chapter gives further insight into the complications of transferring models between native and non-native ranges and is another example of asymmetrical transferability which appears to be a common result in (the currently limited number of ) species distribution model transferability studies. The manuscript for this chapter is co-authored by Tenley M. Conway and Marie-Josée Fortin and will be submitted to *Ecography*. 
Chapter 5

Impacts of climate change and land use change on an invasive species’ distribution

Chapter 5 predicts future areas of suitable habitat for the invasive mute swan by taking into account different scenarios of urban expansion and climate change scenarios. While the correlations between invasive species and urban areas are well known, the relative magnitude of the effects of climate change and urbanization on future suitable habitats of invasive species is not known. This chapter compares predictions of mute swan’s distribution in its non-native range made by two different species distribution models, MaxEnt and a generalized linear model, and discusses possible reasons for variation in predictions. The tangible expected outcomes of this chapter will be risk maps of habitats suitable for mute swan establishment in the future. These will be useful for managers of natural areas and urban parks in terms of preventing or creating management plans for mute swans. The manuscript for this chapter is co-authored by Marie-Josée Fortin and Tenley M. Conway and will be submitted to Conservation Biology.

Chapter 6

Chapter 6 provides a summary of the main results and conclusions, highlights this dissertation’s major contributions to current knowledge gaps and suggests future research directions.
Chapter 2
An Invasive Species’ Relationship with Environmental Variables
Changes when Measured at Different Spatial Scales

2.1. Introduction

An invasive species’ ability to tolerate a wide range of environmental conditions is often one of the characteristics credited with how they can quickly become pervasive outside of their native ranges (Orians, 1986; Sax et al., 2007). Furthermore, there are often fewer significant enemies for a species in recently invaded landscapes, as compared to their native range, further facilitating exotic invasive species establishment and expansion. While many ecologists have focused on the traits of the invader as determinants of invasion success, regional conditions, including climate and landscape spatial heterogeneity, play an important role in determining invasive species initial establishment and subsequent spread (Orians, 1986; Swincer, 1986; With, 2002; Clergeau and Quenot, 2007; Newsome and Noble, 2008). Indeed, exotic invasive species respond differently to environmental variables, possibly at different spatial scales, than do native species, as invasive species often thrive in environments that are highly disturbed and fragmented, such as urban areas (Orians, 1986).

In the context of species distribution modelling, it is important to understand invasive species’ response to the landscape, and how and whether this changes depending on the spatial scale of analysis. The selection of spatial scale is often done by default as a by-product of the data available, and hence species distribution modelling is often based on the spatial resolution of the datasets available, such that the grain (sampling unit size, pixel resolution) and the extent (study area size) are arbitrarily set. However, such set spatial resolutions may not match the underlying ecological processes that influence landscape patterns and species distribution which occur at multiple spatial scales (Mitchell et al., 2001; Schooley, 2006). Ideally, in studying how the given species responds to landscape pattern, all important factors associated with habitat selection at all spatial scales should be taken into account. Indeed, landscape characteristics at broader scales likely reflect climate patterns determining basic habitat suitability and broad resource
availability, while at finer scales, significant characteristics may reflect biological requirements such as food availability and territory characteristics (Wiens, 1989; Hostetler and Holling, 2000). Hence, using only a single-scale model to determine the relationship between species and environmental factors may result in management decisions which do not account for, or incorporate, influential processes occurring at other scales (Hostetler and Holling, 2000; Mitchell et al. 2001; Holland et al., 2004).

One way to address species’ response to environmental variables is to employ a multi-scalar approach (Holland et al., 2004) to measure the effect of different factors at relevant ecological scales for the species of interest. Here, one would essentially measure environmental variables at numerous pre-defined scales which reflect the ecological activities of the species in question. For point count data inputted into a Geographic Information System (GIS), this would involve creating buffers of varying radii around each point. Measuring land cover (i.e. the percentage of each type of land cover) per buffer radii thus essentially scales the independent variables according to relevant scales for the species under study. A single-scale model could be built where all environmental variables are measured at the same scale, or a composite model could be built where each environmental variable may be measured at a different ecologically-relevant scale. A growing number of studies have incorporated a multiple spatial scale framework to determine influential environmental variables on species distribution at territory versus landscape scales (Grand and Cushman, 2003; Blevins and With, 2011; Thornton et al., 2011), with more recent focus on urban areas (Hostettler and Holling, 2000; Pautasso, 2007; Fletcher and Hutto, 2008; Pennington and Blair, 2011). Studies performed at multiple spatial scales can give a more complete understanding of the bird - environment relationships that exist (Pennington and Blair, 2011) given that it is possible that birds may select their habitat based on multiple spatial scales where the most relevant factor may vary between scales (Mitchell et al., 2001; Pennington and Blair, 2011). The scale of landscape measurements will inevitably affect the accuracy and predictive capacity of species distribution modelling, especially of invasive species.

The question therefore remains whether a species’ relationship with the landscape changes across spatial scales (Holland et al., 2004), and what the value of basing management decisions on multiple spatial scales of study is, especially in the context of invasive species. This study
provides evidence as to whether the best-fit models differ between spatial scales of study, and which scale of study is the best for studying invasive species distributions. Here, I use mute swans (*Cygnus olor*) as a case study species to investigate ecologically-based spatial scales, in part because mute swans are large birds with large territories, which lends itself well to studies of landscape correlations from territory to regional scales. Furthermore, there is a need for information about landscape correlations of the mute swan in its non-native range due to current management concerns; mute swans are rapidly increasing in population and have significant effects on submerged aquatic vegetation (hereafter SAV), a food source for multiple aquatic species. Previous studies in Ontario, where the mute swan is considered to be invasive, have focused mainly on the population dynamics of mute swans (Petrie and Francis, 2003). The majority of studies that have considered mute swan habitat preferences have been conducted in their native range of Europe (Holm, 2002; Mason et al., 2006); however, none of these have examined landscape correlations across multiple scales.

Specifically, I address the following questions: (1) Which single-scale model best explains mute swan distribution and at what scale? (2) Are composite models comprised of environmental variables measured at different scales better at explaining mute swan distribution? and (3) What is the relationship between environmental variables and the invasive mute swan’s current distribution and how is this relationship affected by the spatial scale of analysis? By studying the effect of scale on invasive species’ correlations with the landscape, I can determine whether and how using specific spatial scales provides a better understanding of a species’ distribution patterns.

### 2.2. Methods

#### 2.2.1. Study Site

The landscape of southern Ontario is a heterogeneous mix of urban areas, wetlands, forest and agricultural fields (Figure 2-1). It is one of the most densely populated regions in Canada and has experienced a significant loss of wetlands due to rapid urbanization since the early twentieth century (Ducks Unlimited Canada, 2010). The landscape shows a strong latitudinal gradient in which urban land cover is high in far southern Ontario but is replaced with a greater proportion
of wetlands and forest as one moves north. Additionally, many of the larger cities in southern Ontario are found on the larger lakes and waterways. Lakes Ontario and Erie in the south and Lake Huron in the west moderate the region’s climate. There is also an east-west precipitation gradient in the study area.

![Map of study site of southern Ontario (Canada) illustrating urban land cover and mute swan distribution (presence and pseudo-absence data) from 2001-2005 (Cadman et al., 2007; Ontario Ministry of Natural Resources, 2008).]

**Figure 2-1**

2.2.2. Case Study Species: Mute Swans

Mute swans, endemic to Europe and Asia, were likely introduced to Ontario in the early twentieth century, with the first breeding evidence in Ontario found in 1958 (Allin et al., 1987; Ciaranca et al., 1997). They are non-migratory birds, exhibiting short-distance migration in the winter months only if necessary (Kirby et al., 1994). Mute swans are considered to be invasive in
Ontario due to aggressive protection of their territories and nests, their rapid population growth (Petrie and Francis, 2003; Cadman et al., 2007), and their voracious consumption and disturbance of SAV (Kirby et al., 1994). All of these behaviours have affected the aquatic ecosystems in which the swans are located.

The analysis of mute swans and environmental variables are based on presence point data from the 2001-2005 Ontario Breeding Bird Atlas (OBBA) (n=73). The presence point data are a compilation of data collected as point counts (n=36) and a rare bird dataset (n=37). Point counts are obtained by standing at a single location for five minutes and recording all birds seen or heard (Bird Studies Canada et al., 2006). The rare bird dataset is based on a 10km² grid methodology in which all evidence for all breeding birds is recorded, and exact coordinates are recorded when observers happen upon rare bird sightings or breeding evidence (Bird Studies Canada et al., 2006). Data collection took place in summer months and thus results are based on breeding habitats. Captive mute swans were not included in the dataset (Cadman et al., 1987). A drawback of this dataset is that by adding these two datasets, I am combining sampling methodologies and therefore cannot directly account for sampling effort. I believe the doubling in size of the dataset may make this a more robust analysis; however, I also perform an analysis of only the point count data in which the sampling effort is constant.

As generalized linear models require absence data which is often not available with species observations, it is necessary to create a pseudo-absence dataset with the caveat that they are not absolute absences (Barbet-Massin et al. 2012). I created an absence dataset equal in number to the presence points (n=73) (Barbet-Massin et al., 2012). I limited this dataset to the current range of native species with comparable ecology (trumpeter swans (Cygnus buccinator) and Canada geese (Branta canadensis)) to inform the potential maximum extent of mute swan range in southern Ontario (Melles et al., 2003). Absence points were then randomly located within the constrained area in the extent of our study area (Lobo and Tognelli, 2011; Barbet-Massin et al., 2012). Absence points were removed where zones overlapped zones of presence points.
2.2.3. Relevant Ecological Scales

As a species often perceives the landscape at different spatial scales based on their different uses of the land (territory, resources) or biological activities (dispersal, breeding) (Levin, 1992), I conducted the analysis at three perception scales reflecting the extent of these activities as determined by previous mute swan studies (Petrie, 2002; Bowman, 2003; Sousa et al., 2008). The scale of analysis is represented as a perception zone (hereafter ‘ecological zone’) around each presence and absence point (Clergeau and Fourcy, 2005). These ecological zones reflect: (1) mute swans have a relatively small territory size (6.1ha); and (2) based on this territory size, possibly do not disperse very far from their location of birth (3000m); yet (3) males can possibly travel a considerable distance away from their territories during their daily activities (8000m) (Figure 2-2). Thus, the first scale is an ecological zone with a radius of 140m, approximately equal to a circular territory of 6.1ha in size (Petrie, 2002; Petrie and Francis, 2003). The second spatial scale is an ecological zone with a radius of 3000m, which is the result of applying a formula (square root of the area of the territory and multiplying by 12) to the territory size of a species to determine the median natal dispersal distance (Bowman, 2003). This formula was developed by Bowman (2003), who examined the relationship between median dispersal distance and territory size for numerous birds, with the relationship proving very strong for non-migratory species, such as mute swans. Bowman (2003) included mute swans in his analysis and used a territory size of 2.2ha. Here, the Ontario-relevant territory size of 6.1ha is used, resulting in a median dispersal distance of 3000m. Finally, the third spatial scale is 8000m, based on male mute swans’ mean daily travel distance from their territory, however it must be cautioned that this value was determined based on a small sample size (Sousa et al., 2008).
Figure 2-2 Three main ecological scales were used to measure landscape and climate variables to include in the species distribution model; 140m, 3000m and 8000m.

2.2.4. Predictor Variables

Possible environmental variables that could explain mute swan distribution were chosen based on mute swan ecology (Table 2-1). To represent climate, I used two of the BioClim variables (the 1950-2000 average annual mean temperature and average annual precipitation) from the 1km gridded WorldClim dataset (Hijmans et al., 2005). I chose the mean annual temperature and precipitation variables to represent average annual conditions; other temperature variables were highly correlated and therefore were not used. As mentioned previously, mute swans are non-migratory and therefore their territory choice must be sufficient for all seasons (Kirby et al., 1994). These variables were measured at each scale by calculating the mean of each environmental variable over the area of each ecological zone. This data (and the following data) were inputted into a Geographic Information System (GIS) program for further data manipulation and analysis.
Table 2-1 Categories of environmental variables potentially affecting mute swan’s distribution.

<table>
<thead>
<tr>
<th>Landscape</th>
<th>Urban</th>
<th>Climate</th>
<th>Other species</th>
<th>Topography</th>
</tr>
</thead>
<tbody>
<tr>
<td>percent water (%)</td>
<td>road density (m/m²)</td>
<td>mean temperature (°C)</td>
<td>presence of zebra mussels</td>
<td>mean elevation (m)</td>
</tr>
<tr>
<td>percent wetland (%)</td>
<td>percent urban (%)</td>
<td>annual precipitation (mm)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>percent forest (%)</td>
<td>waterbody perimeter density (m/m²)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Land cover data were obtained from the 2000-2002 Southern Ontario Land Resource Information System data (SOLRIS) with 15m resolution (Ontario Ministry of Natural Resources, 2008). The data originally had 25 different land cover types, which I reclassified into five classes (forest, water, wetlands, urban, and other). A benefit of this land cover dataset is its fine resolution (15m) which enabled us to clearly differentiate between urban, water and natural land cover within urban boundaries. As mute swans have been increasingly associated with urban areas in addition to smaller inland waterbodies such as rivers in their native range, this high resolution dataset allows me to more confidently examine landscape correlations at the territory scale. The land cover data were measured per ecological zone as a percent total land cover (i.e. percent forest, percent water, etc.). Road data was obtained from DMTI CanMap Route Logistics (DMTI CanMap Route Logistics, 2011) and the total length of road was measured and then divided by the area of the ecological zone to ascertain road density.

Total waterbody perimeter density was also included as an aquatic environmental variable to reflect potential resource and territory availability in a particular area. The perimeter of the waterbodies was determined using the SOLRIS data described above which was converted from raster to vector data. This variable was measured per ecological zone by summing the length of the perimeter of the waterbody within each zone and dividing by area to produce a density measurement.

Elevation was also included as an environmental variable to reflect possible influences of topography on mute swan distribution (Natural Resources Canada, 2001). The decision to include this environmental variable was informed by the fact that mute swans seem to be rarely
found in elevations greater than 300m in their native range (Kirby et al., 1994). In terms of this study, elevation was available at a resolution of 650m. Mean elevation was calculated for each ecological zone, with the caveat that the elevation at the 140m ecological zone often was only composed of one elevation pixel.

The final environmental variable included was presence of zebra mussels. Zebra mussels are themselves an invasive species that make the water column clearer, reduce algae, and encourage submerged aquatic vegetation growth (Zhu et al., 2006). Thus, presence of zebra mussels may reflect areas with more SAV, the main food source of mute swans. The presence of zebra mussels was expressed as a binary variable, which reflected whether there were zebra mussel sightings within any part of the waterbody, regardless of the waterbody’s size (i.e. marsh or Great Lakes), that overlapped each particular ecological zone in question. These data were obtained from both the Ontario Ministry of Natural Resources and Ontario Federation of Anglers and Hunters (2010) and the United States Geological Survey (2007) who collected this data from 1988 onwards using different data collection methods.

Landscape environmental variables such as bathymetry, reflecting depth of the water, and SAV distribution were not available for all lakes and even though these landscape environmental variables would be relevant for this analysis (Holm, 2002), they could therefore not be included.

2.2.5. Environmental Variable Correlation Analysis

My first step was to calculate Spearman rank correlations to assess environmental variables for multicollinearity. If necessary, I then performed a principal component analysis (PCA) on environmental variables with a correlation value of 0.7 or greater.

Spearman rank correlation analysis revealed that numerous environmental variables were collinear as they had correlation values greater than 0.9. Percent wetland and waterbody perimeter density were highly correlated. As a principal component would be difficult to decipher ecologically, I decided to retain the waterbody perimeter density environmental variable as it was likely to be a more valuable, better reflection of mute swan habitat choice. Percent
water and presence of zebra mussels were also highly correlated; this was likely due to the method used to calculate presence of zebra mussels which meant that larger waterbodies such as Lake Ontario would skew the results as zebra mussel presence anywhere on the waterbody was extrapolated to mean zebra mussel presence at all locations on the waterbody. Due to the unstructured methodologies used to collect zebra mussel data, I decided not to include this environmental variable for this analysis, especially given the nature of the specificity needed for the scale analysis. Finally, the road density and percent urban variables were highly correlated. However, the principal component created for these variables at the 140m zone could only represent 73% of the data, which I deemed to be not enough to include as a single principal component. Preliminary analyses revealed that road density explained mute swan distribution better than the percent urban variable and thus the road density landscape environmental variable was retained.

2.2.6. Alternative Models

2.2.6.1. Single-Scale Models (I-VIII)

Models were created that could best explain mute swan distribution based on my selected environmental variables and knowledge of mute swan ecology (Table 2-2). Model I was a global model that included all of the predictor variables. Models II through VI reflected different groupings of environmental variables that I proposed could best explain certain aspects of mute swan distribution, including presence of urban areas, territory availability, climate, resource availability and land cover. Models VII and VIII reflected combinations of the above models that I proposed could best explain mute swan distribution.
Table 2-2 Eight single-scale and three composite-scale models explaining mute swan’s distribution in its non-native range, and the environmental variables included in each.

<table>
<thead>
<tr>
<th></th>
<th>Percent water (%)</th>
<th>Waterbody perimeter density (m/m²)</th>
<th>Percent forest (%)</th>
<th>Road density (m/m²)</th>
<th>Mean temperature (°C)</th>
<th>Annual precipitation (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Global</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>II. Urban</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>III. Habitat</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV. Climate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V. Resources</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VI. Land cover</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VII. Resources + Climate</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIII. Resources + Urban</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Composite Model 1</td>
<td>8000m</td>
<td>140m</td>
<td>8000m</td>
<td>140m</td>
<td>8000m</td>
<td>8000m</td>
</tr>
<tr>
<td>2. Composite Model 2</td>
<td>8000m</td>
<td>140m</td>
<td>140m</td>
<td>140m</td>
<td>8000m</td>
<td>8000m</td>
</tr>
<tr>
<td>3. Composite Model 3</td>
<td>8000m</td>
<td>140m</td>
<td>8000m</td>
<td>140m</td>
<td>140m</td>
<td>140m</td>
</tr>
</tbody>
</table>

2.2.6.2. Composite-Scale Models (1-3)

I created composite-scale models based on *a priori* hypotheses as to which explanatory landscape environmental variable or combination of environmental variables would be more important at different scales and using the more extreme territory and regional scales. For Composite Model 1, I hypothesized that mute swan distribution is best determined by environmental variables that reflect territory needs at the 140m scale combined with other environmental variables such as climate and land cover being measured at the broader 8000m scale. Thus, for Composite Model 1, waterbody perimeter density and road density were measured at a 140m scale, reflecting possible territory availability and proximity to urban areas, while environmental variables such as percent water and climate were measured at a larger scale, reflecting the broader scale processes such as climate and overall availability of water resources. For Composite Model 2, I hypothesized that in addition to waterbody perimeter density and road
density being measured at the territory scale, the addition of percent forest measured at the 140m scale would possibly better explain the territory selection of mute swans. I hypothesize that forests do not provide a food source for mute swans and likely harbour predators, such as foxes, and as such, mute swans would select territories away from these obstacles (Fuller et al., 2005).

Finally, Composite Model 3 is a modification of Composite Model 2 in which I evaluate whether measuring climate and elevation at a smaller scale has an effect on the suitability of the model.

2.2.7. Statistical Analysis

To test the models, I used generalized linear models with a logit-link function for a binomial dependent variable in R (R Development Core Team, 2012). Environmental variables were assessed for outliers, skewness, normality and heteroscedasticity. Due to the high number of zeros in the predictor variables, especially at the 140m ecological zone, transformations were unable to improve normality of the explanatory variables and hence I used untransformed variables. Yet, the resultant residual plots were normally distributed and overdispersion values were near 1. To enable comparisons between the same single-scale model applied at different ecological zones, all independent variables were standardized to a mean of 0 and a standard deviation of 1 (Smith et al., 2010). To determine the general relationships between each environmental variable and mute swan distribution, and how these may vary across scales, I compared the resulting coefficients of the best single-scale model at each ecological zone.

To determine which model was the best for determining mute swan distribution at a given scale and across scales, we used Akaike’s Information Criterion (AICc) for small sample sizes and the resultant Akaike weights to determine model fit, and area under the receiver operating characteristic curve (AUC) values and Cohen’s Kappa tests for model validation. AICc is a variation of AIC which was developed for datasets in which the ratio of sample size to the number of parameters is less than 40 (Burnham and Anderson, 1998). The lowest AICc values indicate the model with the best fit (Burnham and Anderson, 1998). Average AICc values were determined for each model and then Akaike weights were calculated; Akaike weights are calculated based solely on AICc values and are used to determine the weight of each model (Burnham and Anderson, 1998). Models with higher weights have the best fit. Model predictive accuracy was assessed using AUC values which are predictive accuracy measures independent of
threshold values (Manel et al., 2001). Average AUC values were calculated for each model based on the 100 different splits. Models with AUC values greater than 0.9 were considered to exhibit excellent predictive accuracy, while values of 0.8-0.89 indicated good predictive accuracy and 0.7-0.79 indicated fair predictive accuracy (Swets, 1988). Cohen’s Kappa test measures the amount of agreement between predicted and observed values (Manel et al., 2001). To calculate Kappa values, a threshold must be applied; I applied a prevalence threshold of 0.5 (Manel et al., 2001). Overall, Kappa values of 0.41-0.60 are considered to exhibit moderate agreement, while values of 0.61-0.80 exhibit substantial agreement between predicted and observed values, and values > 0.80 exhibit ‘near perfect’ agreement (Landis and Koch, 1977).

2.3. Results

Overall, when analyzing each scale separately, the global model (Single-Scale Model I), which included all possibly explanatory environmental variables has the highest Akaike weights for each individual scale and thus given our a priori models, it is the best single-scale model explaining mute swan distribution (Tables 2-3 and 2-4). The second best model is Model VII which was based on both resource and climate environmental variables. However, this model still exhibited lower Akaike weights, which means that it has less support than the global models. All other models had Akaike weights of 0.00, indicating that there is little uncertainty in model selection (Burnham and Anderson, 1998). AUC and Cohen’s Kappa values were similar for these two models (Tables 2-5 and 2-6). When the global models for each ecological zone were compared separately, the global model measured at the 3000 m zone has the highest Akaike weight, followed by the global model measured at the 140m zone. In term of the composite models (Table 2-4), all three composite models had higher Akaike weights than the best single-scale global model (notably with Composite Model 2) (Table 2-3) and comparable, if not slightly higher AUC and Cohen’s Kappa values (Tables 2-5 and 2-6).
Table 2-3 Akaike weights of all models at all scales of analysis.

<table>
<thead>
<tr>
<th>Akaike Weights</th>
<th>140m</th>
<th>3000m</th>
<th>8000m</th>
<th>Composite Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Global</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>II. Urban</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>III. Habitat</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>IV. Climate</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>V. Resources</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>VI. Land cover / Topography</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>VII. Resources + Climate</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>VIII. Resources + Urban</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>1. Composite Model 1</td>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>2. Composite Model 2</td>
<td></td>
<td></td>
<td></td>
<td>0.96</td>
</tr>
<tr>
<td>3. Composite Model 3</td>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 2-4 Akaike weights for all single-scale models.

<table>
<thead>
<tr>
<th>Akaike Weights</th>
<th>140m</th>
<th>3000m</th>
<th>8000m</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Global</td>
<td>0.28</td>
<td>0.48</td>
<td>0.00</td>
</tr>
<tr>
<td>II. Urban</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>III. Habitat</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>IV. Climate</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>V. Resources</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>VI. Land cover / Topography</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>VII. Resources + Climate</td>
<td>0.01</td>
<td>0.22</td>
<td>0.00</td>
</tr>
<tr>
<td>VIII. Resources + Urban</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 2-5 Area under the receiver operator curve values for all models with Akaike weights > 0.00.

<table>
<thead>
<tr>
<th>AUC values</th>
<th>140m</th>
<th>3000m</th>
<th>8000m</th>
<th>Composite Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Global</td>
<td>0.896</td>
<td>0.905</td>
<td>0.874</td>
<td></td>
</tr>
<tr>
<td>VII. Resources + Climate</td>
<td>0.884</td>
<td>0.897</td>
<td>0.862</td>
<td></td>
</tr>
<tr>
<td>1. Composite Model 1</td>
<td></td>
<td></td>
<td></td>
<td>0.901</td>
</tr>
<tr>
<td>2. Composite Model 2</td>
<td></td>
<td></td>
<td></td>
<td>0.919</td>
</tr>
<tr>
<td>3. Composite Model 3</td>
<td></td>
<td></td>
<td></td>
<td>0.904</td>
</tr>
</tbody>
</table>
Table 2-6 Kappa values for all models with Akaike weights > 0.00.

<table>
<thead>
<tr>
<th>Kappa values</th>
<th>140m</th>
<th>3000m</th>
<th>8000m</th>
<th>Composite Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Global</td>
<td>0.616</td>
<td>0.641</td>
<td>0.564</td>
<td></td>
</tr>
<tr>
<td>VII. Resources + Climate</td>
<td>0.618</td>
<td>0.615</td>
<td>0.556</td>
<td></td>
</tr>
<tr>
<td>Composite 1</td>
<td></td>
<td></td>
<td></td>
<td>0.630</td>
</tr>
<tr>
<td>Composite 2</td>
<td></td>
<td></td>
<td></td>
<td>0.666</td>
</tr>
<tr>
<td>Composite 3</td>
<td></td>
<td></td>
<td></td>
<td>0.627</td>
</tr>
</tbody>
</table>

Comparing the coefficients of the standardized variables for the global models at each scale revealed that all environmental variables had a positive correlation with mute swan distribution, with the exception of percent forest and mean elevation (Figure 2-3). In addition, when taking into account standard errors, only two factors yielded different levels of influence at different scales. Percent water and annual precipitation both had higher coefficients at the 3000m and 8000m ecological zones. It should also be noted that mean elevation was not a significant variable ($p<0.05$) in the global model for any ecological zone.

![Figure 2-3](image-url) Coefficients and corresponding standard errors of the Global single-scale model applied at the 140m, 3000m and 8000m ecological zones.

2.4. Discussion

Two of my objectives were to test the capacity of single-scale models in predicting mute swan’s distribution and to evaluate the performance of composite models composed of landscape environmental variables measured at different scales reflecting the ecological needs of mute
swans. I chose an information theoretic approach to determining the best models in predicting mute swan distribution, and used AUC values, which measure the accuracy of the predictions versus the observed distributions, to corroborate the results. While the composite models were weighted more heavily when assessed using Akaike weights, I will also address the results and utility of the best single-scale model, as these are often the type of models used in modelling species distributions. In terms of single-scale models, the best model explaining mute swan distribution is the global model measured at a scale of 140m which is reflective of mute swan’s average territory size, and is smaller than the possible cygnet travel distance and the mean travel distance of male mute swans (Petrie, 2002; Bowman, 2003; Sousa et al., 2008).

2.4.1. Composite-Scale Models

The three composite-scale models had higher Akaike weights than the single-scale models measured at any scale, indicating that there is not one perfect scale of analysis for determining mute swan distribution. Other studies have also determined that the best models of species distribution or habitat selection are found when using factors measured at multiple scales (Blevins and With, 2011; Pennington and Blair, 2011; Thornton et al., 2011); though some have found that this leads to little improvement (Mitchell et al., 2001). These three composite-scale models reflected my a priori hypotheses that different processes may have a greater explanatory effect if measured at different scales of analyses. At the territory scale, I expected the most significant environmental variables influencing habitat choice of mute swans to be those related to biology, such as breeding and foraging requirements (Levin, 1992), as opposed to larger-scale physical processes such as climate (Wiens, 1989). These requirements include the availability of sufficient space and resources to build a nest, provision of shelter and the close availability of food sources (Desrochers et al., 2010). I hypothesized that factors affecting territory selection such as percent forest, waterbody perimeter density and road density would have a greater explanatory effect if measured at the territory scale, and that factors affecting regional resource availability such as percent water would have a greater explanatory effect if measured at the broader scales. The negative correlation between percent forest and mute swan distribution is likely more noticeable at the finest scale as the importance of open waterbodies and lack of predators is a localized function and this environmental variable could be overshadowed at the larger scales of analysis. These relationships were found to be true as Composite Model 2 had
the highest Akaike weight, followed by Composite Models 1 and 3, which had less support (Burnham and Anderson, 1998).

2.4.2. The Global Model: Influential Variables

Overall, based on the single-scale global models, I found that the influence of some environmental variables changes with spatial scale, highlighting the value of considering environmental variables at multiple ecological scales. At the broader spatial scales, significant factors determining mute swan distribution likely reflect processes such as intra- and interspecies interactions (competition, conspecific attraction), climate (temperature and precipitation gradients) and resource tracking (Desrochers et al., 2010). These processes are represented by factors such as temperature, precipitation, and percent water. The results of this study are consistent with hypotheses about the processes that are influential at the broader scales, as percent water and annual precipitation had higher coefficient values at the two broadest ecological scales relative to the 140m territory scale.

For all scales of analysis, mute swan distribution was significantly related to percent water. This was expected as mute swans are an aquatic species. Percent water is likely a reflection of the total number of territories allowable in a certain area due to nesting site availability and food availability. In perceiving the landscape, mute swans are likely attracted to larger waterbodies that can provide plentiful food resources and establish at, or in close proximity to these larger waterbodies (Wieloch et al., 2004). Percent water was relatively less important at the 140m ecological zone (i.e. territory size), possibly because percent water reflects resource availability and this may not be the most important factor in territory establishment, where availability of nesting and breeding resources may outweigh immediate proximity of food.

Previous studies have shown that mute swans frequent waterbodies that have SAV, but that these waterbodies vary in proximity to urban land cover (Wieloch et al., 2004). The results of our study indicate that at all ecological spatial scales, mute swan presence is positively related to higher road densities (a proxy for urban land cover), supporting anecdotal evidence that mute swans can thrive in these urban environments. The overall correlation between mute swans and urban land cover is consistent with some studies conducted in mute swans’ native range of
Europe, which found that mute swans are increasingly associated with urban areas (Kirby et al., 1994; Holm, 2002; Mason et al., 2006). However, others have concluded there is no significant correlation (Fuller et al., 2005). None of these previous studies were conducted at multiple scales, nor specifically examined mute swan – landscape correlations.

Mute swans are likely attracted to urban areas due to anthropogenic food sources and habitat opportunities on urban waterfronts. They seem to be able to tolerate the pollution, human traffic and disruptions of urban areas (Holm, 2002). In Ontario, mute swans may also be attracted to urban areas in the winter as urban waterbodies are less likely to freeze due to urban heat island effects and thus can still supply food in the winter. It is interesting that they are associated with urban areas in the summer, the season of this study, when they could likely easily find required resources elsewhere given that they are capable of short-distance migrations (Kirby et al., 1994).

In terms of climate, mute swan distribution is positively correlated with precipitation and temperature at all scales. This translates to areas with a higher average temperature, longer growing season and higher precipitation as more likely preferred habitat. Areas with higher precipitation may also have fewer ephemeral wetlands and perhaps a more consistent and higher quality food supply.

In Ontario, mute swan distribution was found to be negatively correlated with percent forest at the all scales, which is likely a function of mute swans requiring territories in open waterbodies outside forest ecosystems (Wieloch et al., 2004). Mute swans are not known to utilize forest resources, and likely avoid these ecosystems since potential egg and cygnet predators, such as foxes, are found here (Ciaranca et al., 1997; Fuller et al., 2005).

2.5. Conclusion
The field of species distribution modelling is currently advancing at a rapid rate, in which we learn more about the statistical properties and sensitivities of these models through continuing research. In terms of single-scale models, in this study I determined that if one must measure environmental variables at one scale to explain mute swan’s distribution, the best results will be found if measurements are performed using a intermediate scale of 3000m or a finer scale 140m
ecological zone. While most species distribution models are performed at a single scale, the results of our study suggest that composite models allowing measurements to be taken of each factor at a scale that best reflects how that environmental variable is likely perceived by the study species and thus reflecting the ecological needs of the given species may provide models of better fit and similar, if not better predictive accuracy than single-scale models. When analyzing species distributions, I suggest that ecologists consider the scale of the underlying landscape processes and the effect that this may have on their modelling outcomes. As many species datasets are aggregated at the 10×10 km grid scale level, we may need to rethink this methodology for future surveys such that we do not preemptively limit the accuracy of the species distribution models.

When developing species management strategies, the scale at which various ecological activities occur should be considered, as differences in the environmental variables that can explain species distribution at each scale have considerable implications when managing a rapidly growing invasive species population like mute swans in Ontario. The significant positive relationship between mute swan distribution and temperature suggests that any increase in temperature will create more potential habitat for mute swans in Ontario. Thus, management efforts to limit mute swan’s range expansion must account for future climate change scenarios. At the smaller territory scale, this analysis suggests water edges without forest cover are most vulnerable for occupation, making these areas a potential target for nest-deterring management strategies. At all scales of analysis, mute swans select territories located in areas of higher road densities, suggesting that this environmental variable adequately fulfills activity and resource requirements. Thus, management efforts to control these species will be most successful when focused on controlling mute swan populations on urban waterfronts.
Chapter 3

The Effect of Sample Selection Bias on the Predictive Accuracy of Species Distribution Models

3.1. Introduction

Species distribution models are often used for modelling landscape relationships between a species and the landscape conditions, predicting distributions of species in either areas not sampled (Elith and Graham, 2009) or modelling potential species distributions accounting for climate change and land use change scenarios (Araújo et al., 2005; Elith et al., 2010). Improvements have been suggested for species distribution models by studies which have examined the effects of sample size (Hernandez et al., 2006; Wisz et al., 2008a; Bean et al., 2012), sampling bias (Phillips et al., 2009; Bean et al., 2012) and varying range sizes (McPherson et al., 2004; Hernandez et al., 2006; Croci et al., 2007). Specifically, studies have shown that increasing the sample size results in more accurate species distribution model predictions (Wisz et al., 2008a) and that species distribution models are more accurate for species with smaller ecological ranges (Hernandez et al., 2006).

Species distribution models have been used to model current distributions of a species in their native range and have been more recently applied to predicting distributions in invasive ranges (Mau-Crimmins et al., 2006; Fitzpatrick et al., 2007). Few studies have examined; however, whether and how species-landscape relationships may vary within native ranges (Guisan and Theurillat, 2000; Graf et al., 2006; Zanini et al., 2009). Models derived from different regions could result in very different predictions of species distributions when these models are applied to broader areas, even within the native range (Graf et al., 2006).

Spatially, sample selection bias occurs when the training data for a model is drawn from a set of environmental variables which are not independent from the broader extent or data to which they are applied (Phillips, 2008). In many cases, the spatial extent of studies is a pragmatic choice and extent is defined by limited time, effort or data availability, or is simply the entire range of the
species. Sample selection bias differs from transferability as sample selection bias involves using a certain dataset (which could be biased geographically or temporally) to train a model and that model is then applied to a broader area with the same range of environmental variables (Phillips, 2008), whereas transferability would involve applying the model to a geographically distinct area with a different range of environmental variables and potentially different relationships between these variables (see Chapter 4). The implications of using training data that only encompasses a part of the environmental variability found in the study extent has not been extensively studied.

To make accurate predictions of species distributions in either native ranges or their non-native ranges, we need to ensure that the models built are accurately reflecting of the species-landscape relationships. Models predicting invasive species range expansion or location in non-native ranges often rely on models developed using native range presence data. Thus, the ability to create universal and accurate models of the native range is important. To this end, I explore the effects of limiting models in the native range by ecoregions and species density; possible forms of sample selection bias that effectively change the range of the environmental variables used in training the model and the potential utility of the final model when applied to a broader extent. Here, I investigate the following questions: (1) What are the effects of using limited sampling extents (and therefore limited ranges of environmental variables) to model species’ distributions? and (2) Are there general characteristics of sampled landscape that result in more generalizable models?

This study utilizes mute swans as its case study species. Mute swans are native to Europe and Asia, where their distribution covers a broad latitudinal extent and they are often found in heterogeneous environments (Wieloch et al., 2004). This study will focus on the British population, which was estimated to be 25 748 in 1990 (Delany et al., 1992). As a conspicuous, large species with a known habitat association (water), mute swans are easily atlassed. Typically, mute swans are associated with elevations less than 300m (Brown and Brown, 1993; Kirby et al., 1994), and a variety of aquatic habitats including water and wetlands with a preference for rivers and larger wetlands and waterbodies with more room for territories and taxiways for flight (Wieloch et al., 2004; O’Hare et al., 2007). Mute swans rely on submerged aquatic vegetation as their main food source, though they are also often associated with agricultural lands, especially
in Britain (Trump et al., 1994; Chisholm and Spray, 2002). On rivers, they have been found to have a density of one breeding pair per 1.75 km of river though this can be much higher (Trump et al., 1994; O’Hare et al., 2007). Mute swans do not migrate long distances; however, they may migrate shorter distances in the winter to larger waterbodies, which are less likely to freeze and may have better food resources (Holm, 2002).

3.2. Methods

3.2.1. Study area

Generalist species with large continental ranges often do not have complete datasets available that have a consistent sampling methodology, can account for sampling effort or are collected at the same scale. While an atlas is available for avian species in Europe, it is available at a 50km resolution, which is not useful for this study as the mute swan’s relationship with landscape and climate should ideally be studied at a finer scale (see Chapter 1). As such, a complete fine-scale dataset is not available for mute swan’s entire native range (Europe and Asia) and thus I examine the island of Britain, for which I have a complete dataset (Figure 3-1) and posit that, as it covers quite a broad latitudinal extent, with a heterogeneous mix of key climates, topography and land cover, the effects of sample selection bias will be apparent (though the conclusions about implications for the ecology of this species may be limited). This dataset also meets the important assumption necessary for species distribution modelling that the study species is currently in pseudo-equilibrium with the entire extent of the study area (Guisan et al., 2002). As mute swans are a native species in Britain and have had the opportunity to establish in all locations and habitats that they would prefer and time to disperse throughout the landscape (Wieloch et al., 2004), they have likely completed all stages of establishment (dispersal, habitat selection, territory establishment) in Britain.
Britain is a heterogeneous, fragmented landscape with varying land uses dominated by urban, agriculture, grassland and forest cover (Centre for Ecology and Hydrology, 1990). Britain has a broad latitudinal extent and is rich in freshwater ecosystems mainly found in the form of rivers, wetlands and small lakes. Britain has undergone dramatic land use change with the conversion of natural land to agricultural land and with the expansion of urban areas that represented 8.9% of the UK’s total land cover in 2001 (Pointer, 2005).

3.2.2. Species distribution

Mute swan distribution data from the Mute Swan Survey performed in Britain in 1990 was used (British Trust for Ornithology and Wildfowl and Wetlands Trust, 1990). Their methodology divided the countries into 10 km grid cells based on the British National Grid. Surveyors were
given a list of grid cells and told to randomly visit each grid cell and record the number of breeding pairs between April 1 and May 31, 1990 (Delany et al., 1992). A caveat for this dataset was that if surveyors were unable to visit all of the grid cells, they were allowed to make ‘best estimates’ of swan populations (Delany et al., 1992). Due to this procedure, it would be misleading to use the count data but is acceptable to use the presence / absence data per 10 km grid cell. In total, there were 2378 grid cells visited (Figure 3-1).

3.2.3. Absence data

Generalized linear models require both presence and absence data (Lutolf et al., 2006). There are numerous methodologies concerning absence point selection (if using fewer points than in the dataset) or pseudo-absence dataset creation; however, it has been shown that better model results can be found if absence points are located within the same extent as the presence point dataset (VanDerWal et al., 2009). Therefore, a random selection of absence points, equal in number to that of the present points and covering the full extent of the study site were chosen.

3.2.4. Environmental Variables

Land cover and climate variables were included as environmental variables (Table 3-1). Land cover data were obtained from the Corine 1990 Land Cover atlas, which reflect the land cover of Britain at approximately the same time that the mute swan survey was performed (Centre for Ecology and Hydrology, 1990). The data are available in percent land cover per 1 km and has 26 land classifications (Centre for Ecology and Hydrology, 1990). I re-classified these into five: water, wetland, forest, urban and other. I used broad classifications to enable comparisons with land classifications in other regions or non-native ranges. Specifically, mute swans use the shorelines of water bodies and wetlands for habitat in their native range and have been positively associated with urban areas (Wieloch et al., 2004). These environmental variables reflect general land cover preferences for establishing territories and resource-gathering and were measured as a percentage per 10 km grid cell (as determined by the mute swan distribution dataset).
Table 3-1 Environmental variables included in theoretical models to explain mute swan distribution in native range of Britain (with transformations in parentheses).

<table>
<thead>
<tr>
<th>Models</th>
<th>percent urban (log)</th>
<th>percent wetland (log)</th>
<th>percent water (log)</th>
<th>percent forest (log)</th>
<th>mean temperature of warmest quarter (^2)</th>
<th>mean temperature of coldest quarter (^2)</th>
<th>annual precipitation (^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Variables</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Climate</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Land Cover</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Three non-collinear, 1950-2000 averaged climate variables from the WorldClim dataset were used; mean temperature of the warmest quarter, mean temperature of the coldest quarter and annual precipitation (Table 3-1) (Hijmans et al., 2005). The temperature variables represent the more extreme conditions that mute swans must annually endure. These variables were available at a resolution of 1 km and were measured by determining the mean value per 10 km grid cell.

3.2.5. Models

I developed three a priori models which hypothesized which environmental variables would be the most important in determining mute swan distribution (Table 1-1) (Guisan et al., 2002). Three models were compared: (1) all variables; (2) climate, including all temperature and precipitation variables; and (3) land cover, including all land cover variables (percent forest, percent water, percent wetland, percent urban). These models were chosen based on consideration of mute swan ecology, including habitat preferences and landscape correlations (Guisan et al., 2002).

3.2.6. Sub-extents

I first used the base dataset of the entire range of the mute swans in Britain (Figure 3-2a); hereafter referred to as ‘Full Extent’. I then utilized two different methods to divide the native range into two sub-extents. I first divided my study area using the Nature Conservancy’s freshwater ecoregions map (Abell et al., 2008; The Nature Conservancy and World Wildlife Foundation, 2008) (Figure 3-2b). These ecoregions have been applied worldwide and reflect common freshwater fish communities and therefore indirectly reflect water conditions such as
temperature. Britain is divided into two freshwater ecoregions: 402 (Northern British Isles) and 404 (Central and Western Europe). As mute swans are aquatic species, these ecoregions are a reliable method of dividing the landscape into ecologically-based units, possibly reflecting differing mute swan-landscape relationships. Hereafter, these sub-extents will be referred to as ‘Freshwater 402’ (Sub-extent 1) and ‘Freshwater 404’ (Sub-extent 2).
Figure 3-2 Sub-extents used for geographically dividing full extent (a) of study site. These were divided using freshwater regions – Freshwater 402 and Freshwater 404 (b), and mute swan presence density – Outer Core and Inner Core (c).
I also created sub-extents based on presence density (Figure 3-2c). The purpose of this division was to analyze whether models created based on a high presence-point density significantly differed from those models created from a low presence-point density (Peterson et al., 2007). This examined the potential effects of when studies may use biased presence data, which are collected only where high densities of that species are located. While sampling effort was not available to control for sampling biases in this dataset, this problem is somewhat ameliorated by the fact that we have aggregated this data as presence and absence data, and that grid cells that were not visited were not included in the analysis. Even if this is slightly biased, I posit that the range of environmental variables of the dataset based on areas of high mute swan density will be reflective of mute swan’s ideal land cover and climate conditions.

For this analysis, I converted the presence grid cells for the entire study area to centroid points. I then estimated the kernel density and created 50% isopleths based on these estimates. I used the least squares cross-validation (LSCV), smoothed cross-validation (SCV), and Plug-In and solve-the-equation methods of determining the optimal bandwidth (Gitzen et al., 2006). The SCV and Plug-In methods produced similar home ranges and ultimately the SCV home range was used. Two sub-extents were thus created based on an inner core encompassing 50% of the presence points and a second sub-extent including all remaining presence points (outer core); hereafter referred to as the ‘Outer Core’ (Sub-extent 3) and the ‘Inner Core’ (Sub-extent 4).

3.2.7. Datasets

For consistency, I developed a dataset with equal numbers of presence (1037) and absence (1037) cells (Figure 3-1). I randomly partitioned the entire presence / absence (1037 / 1037) dataset ten times with 70% for training and 30% for validation (Figure 3-3) to initially create ten ‘mega-datasets’. Thus, 30% of the original sample was always used for validation and 70% for model training (Randin et al., 2006). For each sub-extent, I then took the training data for each ‘mega dataset’ and randomly selected 150 presence points and 150 absence points ten times (i.e. ten replicates). As the smallest region had a total of approximately 170 presence points (though this varied by dataset), I chose 150 presence points as a reasonable number to use in each region (Peterson et al., 2007). The sub-sampling of the data within sub-extents helped to account for any sample biases and created equal prevalence within sub-extents (Araújo and Guisan, 2006).
As models with larger sample sizes will inherently have higher accuracy as shown by higher area under the receiver operator characteristic curve (AUC) values (McPherson et al., 2004), I maintained an equal sampling dataset size for each sub-extent. As I had ten ‘mega datasets’ to begin with; this resulted in 100 different datasets to test each sub-extent.

![Flow chart of data partitioning and dataset creation for each sub-extent.](Image)

**Figure 3-3** Flow chart of data partitioning and dataset creation for each sub-extent.

### 3.2.8. Statistical Analyses

To determine which environmental variables were highly correlated, I performed Spearman rank correlations using the original presence / absence dataset (1037/1037). Only variables which were not collinear were included. The distribution of each variable was assessed for normality and log transformations were performed where necessary. For each individual dataset, I then performed a logistic regression using a binomial generalized linear model with a logit link. Models were evaluated for possible overdispersion and spatial autocorrelation of the residuals.
3.2.9. Model comparison

For every run of every model of each dataset, Akaike Information Criterion for small sample sizes (AIC<sub>c</sub>) was calculated. AIC<sub>c</sub> is an evaluation of the fit of a model, and effectively penalizes for the number of parameters used with lower AIC values indicating better fit (Rushton et al., 2004). AIC<sub>c</sub> is a variation of AIC and is applicable when the division of the sampling size by the number of environmental variables is less than 40 (Burnham and Anderson, 1998). The mean and standard deviation of the AIC<sub>c</sub> values for each sub-extent were calculated and Akaike weights were calculated based on the mean AIC<sub>c</sub> values for each sub-extent. The best model (as selected from the ‘all variables’, ‘climate’ and ‘land cover’ models) for each region was chosen based on highest Akaike weight.

3.2.10. Measuring predictive accuracy

For each of the best sub-extent models, one run of the model was used to represent each sub-extent for the final comparison. This run was chosen by determining the run with the highest Nagelkerke $R^2$ value, and therefore explained the most variation in the model for that sub-extent. Nagelkerke’s $R^2$ values (a pseudo $R^2$ estimate for logistic models) are a measure of how much variance a model explains (Graf et al., 2006). Like $R^2$ values, the possible range of values is between 0 and 1.

The value for the area under the receiver operating characteristic curve (AUC) was determined for each of these predictions. AUC is a measure of the ability of the model to correctly predict both presence and absence points and accounts for incorrect predictions (false negatives and positives) (Fawcett, 2006). To evaluate sensitivity and kappa values, I also calculated a presence threshold for each sub-extent based on the minimized difference between sensitivity and specificity values (MDSS Threshold) determined by the sub-extent in which the model was built (Liu et al., 2005). This method of determining thresholds has been established as being more accurate than fixed thresholds (Liu et al., 2005) and adjusts for species prevalence based on the area in which the models were built.

To evaluate the percentage of the presence dataset points correctly predicted, sensitivity values of the entire study extent were calculated using predictions based on models built in each sub-
extent (McPherson et al., 2004). Sensitivity values differ from AUC values as sensitivity values are only concerned with the accurate prediction of presence points. While AUC values account for correct and incorrect predictions of presence and absence, I cannot be as confident in the accuracy of the absence cells as I can be with the presence cells (I can know for sure the mute swan is there but cannot be sure that the mute swan is not there) and therefore I was more interested in determining how accurate the best models were in predicting known presence cells. The AUC and sensitivity values for each sub-extent were calculated based on the validation datasets, as these extrinsic measures of model accuracy are much more useful than those based on intrinsic measures (training data) (McPherson et al., 2004; Araújo et al., 2005). The qualitative analysis of the AUC values were based on Swets (1988) who characterized AUC values of 0.7-0.79 as fair, 0.8-0.89 as good and 0.9-1.0 as excellent.

3.2.11. Range of Environmental Variables
To ascertain the differences between land cover and climate characteristics of each of the sub-extents, mean values and standard deviations were calculated for each environmental variable in each sub-extent. Histograms of each environmental variable were also created (using equal sample sizes for all sub-extents).

3.2.12. Variable Importance
Variable importance was determined by calculating Nagelkerke’s $R^2$ values to estimate the amount of variance explained by each individual variable. Nagelkerke’s $R^2$ values were determined for the full model for each sub-extent, and these values were again determined for a reduced model whereby each environmental variable was removed individually. The percentage change in the $R^2$ value for the model was determined for each environmental variable and this was used as an estimate of the variable importance. An important note with this methodology is that it does not account for correlations between variables and therefore individual contributions of each variable must be examined with this caveat in mind.
3.3. Results

3.3.1. Comparison of models

The ‘All variables’ model including all environmental variables was the best model for explaining the four sub-extents and the full extent (Table 3-2). These models had the highest Akaike weight values, indicating less support for the other models per sub-extent (Burnham and Anderson, 1998). However, it should be noted that there is some model uncertainty as any models with an Akaike weight value greater than 0.7 were used as the sole model to explain the distribution in that sub-extent.

Table 3-2 Akaike weights (evaluated within sub-extents) for each model built at seven spatial sub-extents.

<table>
<thead>
<tr>
<th>Model</th>
<th>All Variables</th>
<th>Climate</th>
<th>Land Cover</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial Extent</td>
<td>Akaike Weights</td>
<td>Akaike Weights</td>
<td>Akaike Weights</td>
</tr>
<tr>
<td>Full Extent</td>
<td>0.93</td>
<td>0.00</td>
<td>0.07</td>
</tr>
<tr>
<td>Sub-extent 1: Freshwater 402</td>
<td>0.99</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>Sub-extent 2: Freshwater 404</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Sub-extent 3: Outer Core</td>
<td>0.73</td>
<td>0.00</td>
<td>0.26</td>
</tr>
<tr>
<td>Sub-extent 4: Inner Core</td>
<td>0.97</td>
<td>0.01</td>
<td>0.02</td>
</tr>
</tbody>
</table>

3.3.2. Predictive accuracy of models

The predictive accuracy of these models when applied to the entire study extent varied between sub-extent models, with the Freshwater 404 and Inner core sub-extents of the aforementioned sub-extents having the lowest sensitivity values (Table 3-3). The general trend was that the models for the Full Extent, Freshwater 402 and Outer Core sub-extents were the models best able to predict mute swan presence over the entire study extent as reflected by the highest sensitivity and Kappa values. Models based on Freshwater 402 and Outer core sub-extents correctly predicted 87% and 83% respectively of mute swan presence cells in the validation data for the entire study extent. The models based on the Full Extent performed similarly and was able to correctly predict 77% of mute swan presence cells in the validation data for the entire study.
extent. In contrast, the Freshwater 404 and Inner Core sub-extents had lower sensitivity values and were only able to correctly predict 57% and 58% of mute swan presence cells respectively.

Table 3-3 AUC values of the best models of the best run per sub-extent and AUC, Kappa and sensitivity values based on application of each model to the entire study extent.

<table>
<thead>
<tr>
<th>Spatial Extent</th>
<th>Sub-Extent</th>
<th>AUC</th>
<th>Study Extent</th>
<th>AUC</th>
<th>Kappa</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full Extent</td>
<td>0.79</td>
<td></td>
<td>0.79</td>
<td>0.43</td>
<td>0.77</td>
</tr>
<tr>
<td>Freshwater 402</td>
<td>Sub-extent 1:</td>
<td>0.79</td>
<td></td>
<td>0.77</td>
<td>0.38</td>
<td>0.87</td>
</tr>
<tr>
<td>Freshwater 404</td>
<td>Sub-extent 2:</td>
<td>0.77</td>
<td></td>
<td>0.78</td>
<td>0.39</td>
<td>0.57</td>
</tr>
<tr>
<td>Outer Core</td>
<td>Sub-extent 3:</td>
<td>0.72</td>
<td></td>
<td>0.78</td>
<td>0.41</td>
<td>0.83</td>
</tr>
<tr>
<td>Inner Core</td>
<td>Sub-extent 4:</td>
<td>0.72</td>
<td></td>
<td>0.73</td>
<td>0.32</td>
<td>0.58</td>
</tr>
</tbody>
</table>

3.3.3. Range of environmental variables
A trend can also be seen in the ranges covered by each environmental variable when comparing sub-extents and the full extent (Figure 3-3). The two suites of environmental variables are the same as previously mentioned, with the Full Extent, Freshwater 402 and Outer Core sub-extents all having broader ranges of values per environmental variable while the Freshwater 404 and Inner Core sub-extents have more narrow ranges of values of each predictor variable.
Within each sub-extent, the distribution of the environmental variables also exhibited different patterns. The following four histograms illustrate the distribution of the four most important variables as determined by the Nagelkerke’s $R^2$ values: percent urban (Figure 3-5), annual precipitation (Figure 3-6), mean temperature of warmest quarter (Figure 3-7) and mean temperature of coolest quarter (Figure 3-8). Most notable in terms of the percent urban land cover histogram is that the Freshwater 404 and Inner Core sub-extents have fewer grid cells with 0% urban land cover than the other sub-extents do; indicating that percent urban land cover is higher in these sub-extents (Figure 3-5).

**Figure 3-4** Mean and range of the values of each environmental variable used per sub-extent.
In terms of annual precipitation, there is a notable trend in the sub-extents with the Freshwater 404 and Inner Core sub-extents exhibiting a peak in their values of annual precipitation at a much lower value than where the Freshwater 402 and Outer Core sub-extents’ values peak. The Full Extent exhibits more of an average trend in terms of annual precipitation.
Temperature histograms also notably vary between the Freshwater 404 and Inner Core sub-extents when compared with the Freshwater 402 and Outer Core sub-extents. The former exhibit higher mean temperatures of the warmest quarter and less variation in the mean temperatures of the coolest quarter when compared to the Freshwater 402 and Outer Core sub-extents.

**Figure 3-7** Histogram of mean temperature of warmest quarter variable for all four sub-extents and full extent.

![Histogram of mean temperature of warmest quarter variable](image)

**Figure 3-8** Histogram of mean temperature of coolest quarter variable for all four sub-extents and full extent.

![Histogram of mean temperature of coolest quarter variable](image)
3.3.4. Variable importance

For each sub-extent and the full extent, the percent urban land cover, percent water, temperature and precipitation variables were the most important variables in terms of their contribution to the models, and evidenced by higher Nagelkerke’s $R^2$ values (Table 3-4). The variable with the highest importance differed between sub-extents: the percent urban variable was most important for the Freshwater 402 sub-extent, the percent water variable was the most important for the Freshwater 404 sub-extent, and the temperature variables were the most important variables for the Outer and Inner Core sub-extents and the Full Extent. Nagelkerke’s $R^2$ values varied between sub-extents and were highest for the Freshwater 402 sub-extent (0.57), followed by the Full Extent (0.56), Freshwater 404 sub-extent (0.53), Outer Core sub-extent (0.46) and Inner Core sub-extent (0.42).

<table>
<thead>
<tr>
<th>Environmental Variables</th>
<th>Percent Loss of Explained Deviance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full Extent</td>
</tr>
<tr>
<td>Percent Urban (%)</td>
<td>4.32</td>
</tr>
<tr>
<td>Percent Forest (%)</td>
<td>4.14</td>
</tr>
<tr>
<td>Percent Wetland (%)</td>
<td>0.00</td>
</tr>
<tr>
<td>Percent Water (%)</td>
<td>4.32</td>
</tr>
<tr>
<td>Mean Temperature of Warmest Quarter ($^\circ$C*10)</td>
<td>0.90</td>
</tr>
<tr>
<td>Mean Temperature of Coolest Quarter ($^\circ$C*10)</td>
<td>7.93</td>
</tr>
<tr>
<td>Annual Precipitation (mm)</td>
<td>5.95</td>
</tr>
</tbody>
</table>

3.4. Discussion

The investigation into the effects of using subsets of data to model species’ distributions outside their areas of calibration is an important aspect in species distribution modelling and is mainly a function of availability and reliability of data (Peterson et al., 2007; Phillips, 2008). The impacts...
of sample selection bias on thresholds, distribution predictions and the performance of species distribution models has recently garnered more attention (Phillips, 2008; Phillips et al., 2009; Bean et al., 2011; Lobo and Tognelli, 2011). Sample selection bias can result in models that do not account for all of the variability found in the distribution of a particular species, or for marginal or extreme habitats (Phillips, 2008).

3.4.1. Models – all variables vs. climate vs. land cover

Overall, the ‘All variables’ models exhibited better goodness-of-fit than the other two models as indicated by higher Akaike weights (Table 3-2). This is likely due to the latitudinal extent and heterogeneity of the landscape that those models using only a limited number of environmental variables could not explain mute swan distribution as well as the model including all predictor variables. However, one ‘land cover’ model had an Akaike weight greater than 0.1, indicating that the this model also had some support in explaining species distributions (Burnham and Anderson, 1998).

3.4.2. Predictive accuracy

While each sub-extent may have similar predictive accuracy values (indicated by comparable AUC values) when assessing predictions for the entire study extent, the differences in sensitivity values indicate that there are differences in each sub-extent’s ability to predict mute swan presence cells in the validation datasets for the entire study extent. As the AUC values account for both commission and omission errors and can be misleading (Lobo et al., 2007), I believe that in this study, the sensitivity values provide more valuable information as they measure the ability of each sub-extent’s model to predict presence. In addition, determination of the MDSS threshold at which to measure sensitivity values ensures that these performance measures are not misleading by over-predicting presence cells. As Graf et al.’s (2006) study on capercaillies corroborated, some of these models were able to explain the variation in their respective sub-extents very well; however, they did not necessarily maintain their high values of predictive accuracy when applied to the entire study extent. This is likely because some models are overfitted to the landscape and climate conditions found in their respective sub-extents and do not account for the heterogeneity and variability in environmental variable values found over the entire study extent (Graf et al., 2006; Randin et al., 2006; Austin, 2007). Furthermore, the
importance of the variables in that sub-extent may not be reflective of the broader relationships and variable importance for the entire extent.

Lower sensitivity values were found for models derived from the Freshwater 404 and Inner Core sub-extents, while higher sensitivity values were found for the models derived from the Freshwater 402 and Outer Core sub-extents in addition to models derived from the Full Extent. This partition between sub-extents in terms of the patterns seen continues throughout my analysis in terms of both predictive accuracy, ranges of environmental variables and general sub-extent characteristics.

3.4.3. Environmental variables

The key to explaining the differences in the abilities of the different sub-extent models to accurately predict mute swan presence throughout the entire study extent lies in the range of values of each environmental variable included in each sub-extent’s training data in addition to the general patterns illustrated by the histograms of the most important variables.

Lower sensitivity values were found for models derived from the Freshwater 404 and Inner Core sub-extents, which reflected a more highly urbanized land cover and less climate variability. This poor performance can be generally attributed to the inclusion of mainly optimal habitats and thus relationships at the edge of the range (less optimal locations) are not predicted as well. Comparison of the ranges of environmental variable values of these sub-extents with the aforementioned Freshwater 402 and Outer Core sub-extents and the Full Extent further illustrates these variations, especially when examining the climate variables (Figure 3-4).

The histograms of the Freshwater 404 and Inner Core sub-extents’ variables (Figures 3-5, 3-6, 3-7, and 3-8) generally illustrate narrow ranges of environmental conditions, as demonstrated by the narrow, high peaks in the histogram around mean values. For the Inner Core sub-extent (derived from the region with the highest mute swan density), this likely reflects the optimal conditions of the mute swan in Britain. In contrast, models derived from regions with more heterogeneous environments, that likely have more extreme environments (more varied topography, colder climates) and generally broader ranges of environmental variable values,
exhibit better predictive accuracy because they are better able to account for heterogeneous environments (Graf et al., 2006; Randin et al., 2006). Examples of these types of models include the models derived from the Full Extent, and the Freshwater 402 and Outer Core sub-extents. These sub-extents cover largely heterogeneous land cover types, which are not dominated by urban land cover, and have broader ranges of values for each environmental variable. This pattern is noticeable especially for percent wetland, percent water, precipitation and temperature variables. The three aforementioned sub-extents are also characterized by higher mean annual precipitation and lower mean annual average temperatures.

While the two suites of sub-extents (the first including the Freshwater 404 and Inner Core sub-extents versus the Freshwater 402 and Outer Core sub-extents in addition to the Full Extent) have the aforementioned general characteristics, comparing the loss of explained variance when removing individual variables from each sub-extent’s model sheds little light on why the Freshwater 402 and Outer Core models are able to perform better than the Full Extent model when predicting mute swan presences throughout the study extent. For example, the Freshwater 402 model has its greatest loss of explanation of variance when the percent urban variable is removed; however, this variable is not as important in either the Full Extent or Outer Core models. It is likely the model’s incorporation of complex interactions between the variables that is resulting in the Freshwater 402 model being able to account for the heterogeneity of the study extent without over-predicting presence points.

3.5. Conclusions

Models built in regions that account for more variability and range in environmental variables and marginal habitats have higher predictive accuracy of predicting species distributions when applied to the broader regions. These models account for heterogeneity of the landscape and climate relationships and therefore can have higher accuracy in predicting species distributions in these and other regions. These results can be generalized to other generalist species with broad ranges in habitat requirements.

However, it is noteworthy that models based on more homogenous sub-extents with less marginal habitat and range of each environmental variable can have lower predictive accuracy,
particularly of presence cells, when applied to other regions. This is particularly concerning when studies do not account for the source and extent of their data, and the potential implications this may have on the accuracy of their predictions.

I therefore recommend that any species distribution model research that uses models developed in one range and applies these to another consider the actual landscape and climate characteristics of these areas, and especially the level of heterogeneity accounted for in the original data used to build the models. In terms of application to species with expanding ranges, models that are developed based on a large portion of the native range and potentially the edges of the current range are better able to model species-landscape relationships. Models based on areas of high species presence or potentially areas with one dominant land cover may not reflect the true extent of species-landscape relationships. It is ideal to use the sampling data for the entire native range (and I recommend further research doing this with both mute swans and other species). However, when making predictions based on data from smaller regions within a range or when extrapolation a model beyond the edges of the current range, I conclude that, as others have suggested (Araújo et al., 2005; Randin et al., 2006), we must exercise caution and define the limitations of the data used in building the models and the predictions that can be made from these.
Chapter 4
Assessing Model Transferability:
Modelling Recently Established Invasive Species
Distributions in Native and Non-Native Ranges

4.1. Introduction

Species distribution models are commonly used for establishing species-landscape relationships through correlative studies, using methods such as classification trees (De’Ath and Fabricus, 2000; Cutler et al., 2007; Capinha and Anastacio, 2011), or for predictions of range expansion, using methods such as generalized linear models (GLMs), machine learning models or genetic algorithm for rule-set production (GARP) (Araújo et al., 2005; Munoz and Real, 2006; Randin et al., 2006). More recently, studies have applied species distribution models to reciprocally model invasive species in their native and non-native ranges (Broennimann et al., 2007; Fitzpatrick et al., 2007). This effort is complicated as invasive species are often expanding their ranges and may not have fully established in their non-native ranges and are therefore not in equilibrium with the environment, a key assumption of species distribution models (Elith et al., 2010; Václavik and Meentemeyer, 2012).

The ability to apply a model built in one region with certain environmental characteristics to another with different environmental characteristics is termed ‘transferability’ (Randin et al., 2006; Phillips, 2008). A key aspect of building models that are transferable is that generality may be more important than precision (Wenger and Olden, 2012); an objective that can be contrary to general species distribution model building preferences (Graf et al., 2006). As such, there has been limited success in creating transferable models thus far (Randin et al., 2006; Fitzpatrick et al., 2007; Peterson et al., 2007; Zanini et al., 2009), and most studies have recommended caution (Randin et al., 2006; Phillips, 2008). True transferability is assessed by building models in two (or more) different geographical ranges and assessing whether they both show good predictive accuracy when applied to the reciprocal range.
Two key assumptions are made when transferring a model involving invasive species: (1) that the species-climate-landscape relationships have not significantly changed between native and non-native environments; and (2) that the species in the range you want to predict is also in equilibrium with the environment (Elith et al., 2010). The first assumption presumes that there is little biotic change in a species’ relationship with the environment between its native and non-native range. This is unlikely as biotic interactions, competitors and resource availability (and the changing relationships between these) are likely to vary even within native ranges, let alone between ranges (Randin et al., 2006; Zanini et al., 2009).

The second assumption posits that the species is found in all ideal habitats (Elith et al., 2010). However, in non-native ranges, the species may not be at equilibrium with the environment, especially if it has been recently introduced (Václavik and Meentemeyer, 2012). Violations of these two assumptions can pose a problem for species distribution models as they assume that species are all found where they have suitable habitat and are not found where they do not have suitable habitat, based on stable habitat preferences.

In this study, I address the following question: Can a species distribution model for the invasive mute swan be successfully transferred between parts of its native and non-native ranges? For my case study species, I hypothesize that using models constructed in the native range to predict distribution in the non-native range will likely result in an over-prediction of the current mute swan distribution in the non-native range. This is because mute swans have relatively recently established in Ontario (1958) and anecdotally are unlikely to have dispersed to all possible habitats within Ontario yet. Thus, the range of habitats which they currently occupy in their non-native range may be limited in comparison to that found in the native range. Successful transfer of these species distribution models would indicate that the landscape relationship is stable between the two ranges, which will provide useful information for managing this invasive species. If the landscape relationship differs, this will provide insight into the characteristics of landscapes susceptible to invasion by the given species. Within my broader research question, my objectives are also to determine whether the most important environmental variables related to mute swan distribution in the native range are consistent with those environmental variables
found to be important in the non-native range, and whether certain environmental variables (e.g. elevation) should be included in predicting distributions when predicting outside of the calibrating range. Determining the most important landscape variables correlated with mute swan distribution will help to ascertain where the mute swan is likely to spread next as the population grows in its non-native range, as mute swans are not yet established in all possible areas of Ontario due to their relatively recent arrival.

As mentioned previously, I utilize mute swans (*Cygnus olor*) as my case study species. Mute swans are non-migratory birds endemic to Europe and Asia and are found throughout Britain (Allin *et al*., 1987), but are considered to be an invasive bird in North America. In 1990, the British population of mute swans was estimated at 25,748 but it has steadily increased to approximately 31,700 in 2002 (Delany *et al*., 1992; Ward *et al*., 2007). Mute swans were first introduced to North America via the urban parks of New York City in the late 19th century (Ciaranca *et al*., 1997).

Since the late 19th century, mute swans have been introduced to other locations throughout North America and have escaped or have purposefully been released into the wild. In Ontario, the mute swan was first sighted in 1934, with the first evidence of breeding in the wild found in 1958 (Ciaranca *et al*., 1997). In 2005, the non-captive Ontario population was estimated to be 2,737 birds, which was almost double the estimated population in 2002 (Cadman *et al*., 2007). Mute swans are considered invasive in Ontario due to their rapid population growth and competition for food and habitat with native waterfowl (Petrie and Francis, 2003).

### 4.2. Methods

#### 4.2.1. Study Sites

The British landscape is a heterogeneous mix of urban areas, wetlands, forest and agricultural fields. There is a north-south temperature and population density gradient with both variables decreasing with increase latitude, although there are noticeable population hotspots throughout the country. Overall, Britain is highly urbanized; in 2001, 9% of the United Kingdom’s (Britain and Ireland) land mass was composed of urban areas (Pointer, 2005).
The landscape of southern Ontario has a similar mix of land uses, and is one of the most densely populated regions in Canada. Southern Ontario also exhibits a latitudinal land cover gradient, with a decrease in urban areas, an increase in the number and size of wetlands, and an increase in forest patch size with increasing latitude. Much of the urban development in Ontario has taken place around water bodies. Within the context of increasing urbanization and loss of wetlands, the invasive mute swan is thriving in its non-native range.

In addition to similar climatic and population gradients, Britain and Ontario are both moderated by nearby large bodies of water: the North Sea and British Channel in Britain and the Great Lakes in southern Ontario.

4.2.2. Mute Swan Distribution Datasets
Mute swan data for Britain were obtained from the National Mute Swan Survey performed in 1990 (British Trust for Ornithology and Wildfowl and Wetlands Trust, 1990). The landscape was surveyed using 10×10 km grid squares in which observers were instructed to visit all possible wetland habitats during the breeding season (April 1 – May 31) and record all territorial and non-breeding birds (Delany et al., 1992). Overall, 85% of the squares were visited by volunteers (Delany et al., 1992). Abundance counts are given for each 10×10 km square, which I transformed into a presence and absence dataset (British Trust for Ornithology and Wildfowl and Wetlands Trust, 1990).

The mute swan data in Ontario are based on presence data from the 2001-2005 Ontario Breeding Bird Atlas (OBBA) (Bird Studies Canada et al., 2006). The Ontario Atlas’ approximate 10×10 km grid squares dataset reports highest breeding evidence found by volunteers in each square (Bird Studies Canada et al., 2006). Volunteers are instructed to visit a variety of sites within the 10×10 km grid square in the breeding months and data from the 2001-2005 visits are compiled (Bird Studies Canada et al., 2006). It is prudent to note that the OBBA makes a concerted effort to exclude all captive mute swans from the datasets, such that all data points represent wild birds (Cadman et al., 1987).
Generalized linear models have been used extensively to model species distributions (Randin et al., 2006; Broennimann et al., 2007; Zanini et al., 2009; Elith et al., 2010). To utilize generalized linear models, presence and absence data are required. Yet generalized linear models can be strongly affected by datasets which contain a highly skewed prevalence. Hence, I utilized an equal number of presence and absence points for each dataset (Britain and Ontario). For Britain, the extent of the possible absence dataset was confined to the extent of the mute swan presence dataset and an equal number (1037) of grid cells were chosen from the 1341 grid cells where mute swans were not found in the survey (Delany et al., 1992). In southern Ontario, the mute swan is still expanding its range; thus, a refined absence dataset for Ontario was selected based on the current range of native species with similar habitat requirements to determine feasible locations where mute swans could disperse to but have not yet established (Melles et al., 2010). Specifically, potential absence cells were confined within the current range extent of trumpeter swans and Canada geese (Cadman et al., 2007). A refined absence dataset was used to control the prevalence of the sampling data to 50%.

4.2.3. Predictor Variables
Climate and landscape environmental variables affect habitat requirements for bird activities such as breeding, roosting and feeding. I included climate and landscape environmental variables that were associated with mute swan distribution based on previous studies (Kirby et al., 1994; Wieloch et al., 2004, see Chapters 2 and 3). While other more specific variables could have been included, such as soil type, I wanted to preserve the generality of the model as much as possible for comparison between the native and non-native ranges.

For mute swan’s native range of Britain, the Land Cover Map of Great Britain (1990) was used to compute the percent land cover per 10×10 km grid cell (Centre for Ecology and Hydrology, 1990). This Land Cover Map consists of 26 different land cover types that I reclassified into forest, water, wetlands, urban and other. I separated aquatic variables into water and wetlands to more distinctly represent mute swan’s purported aquatic habitat. For Ontario, the 2000-2002 Southern Ontario Land Resource Information System (SOLRIS) data consist of 25 different land
cover types at a resolution of 15 m, which I reclassified to align with the British data (Ontario Ministry of Natural Resources, 2008).

Climate and elevation data for both Britain and Ontario were obtained from the WorldClim database (Hijmans et al., 2005). Initially, mean average 1950-2000 temperature, diurnal temperature range, mean maximum temperature per warmest quarter, mean minimum temperature per coldest quarter, mean annual precipitation, mean maximum precipitation per wettest quarter and mean minimum precipitation per driest quarter were selected. These variables were chosen to represent both average conditions and the extreme temperatures and precipitation that may affect mute swan distribution in the landscape. Given that mute swans are non-migratory, mute swans may be more influenced by extreme temperatures than average values (Kirby et al., 1994). However, Spearman Rank correlations indicated high collinearity between the respective temperature and precipitation variables so only two variables were used: mean temperature and mean annual precipitation. The elevation variable was included as previous studies in mute swan’s native range have shown that mute swans are generally associated with habitats less than 300 m in elevation (Kirby et al., 1994). I also wanted to examine the use of elevation as an environmental variable as its use in species distribution models has been questioned (Austin, 2007). Austin (2007) proposed that the relationship between elevation and temperature and precipitation variables is a function of local conditions and these relationships are unlikely to be consistent when transferred to other areas. The average values for the climate and elevation variables were calculated for each 10×10 km grid cell to conform to the resolution of the mute swan data, both in Britain and Ontario.

4.2.4. Models

Four models were built to test mute swan-landscape relationships (Table 4-1). The first (Model I – Land cover + Climate + Elevation (hereafter referred to as LCCE)) included all seven predictor variables. Model II (Land cover + Climate (hereafter referred to as LCC)) is a variation of Model I with elevation removed to test the effect of including elevation. Model III (Habitat and Climate) reflects only those variables which are necessary for mute swan survival, including access to water, warmer mean elevations, areas with higher precipitation and access to urban
areas, as mute swans have been found to have a high correlation with these areas, perhaps due to their utility for resources, lack of competition and urban heat island effect (Kirby et al., 1994). Finally, Model IV (Land Cover) only includes land cover environmental variables that may be more dominant in mute swan habitat preferences than climatic conditions.

**Table 4-1** Theoretical models explaining mute swan (*Cygnus olor*) distribution in its native (Britain) and non-native (Ontario) ranges (with transformations of environmental variables in parentheses).

<table>
<thead>
<tr>
<th>Theoretical Models</th>
<th>Percent water (log+0.1)</th>
<th>Percent wetland (log+0.1)</th>
<th>Percent forest (log+0.1)</th>
<th>Percent urban (log+0.1)</th>
<th>Mean annual temperature (^2)</th>
<th>Annual precipitation (^2)</th>
<th>Elevation (log+10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Land Cover + Climate + Elevation (LCCE)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Land Cover + Climate (LCC)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Habitat + Climate</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Land Cover</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**4.2.5. Statistical Analysis**

**4.2.5.1. Dataset Creation**

Britain’s 1037 presence and 1037 absence cells were partitioned one hundred times into 70% training and 30% validation sets. Thus, every iteration was tested on data not used to calibrate the model. I did this to test if a reasonable model could be built and validated in the calibrating range before applying it to the other range.

In Ontario, I used the same partitioning, with 134 presences and 134 randomly selected (from 997 potential) absence grid cells. Even though mute swans are likely not in equilibrium in their non-native range, they have had the opportunity to establish in more ideal habitats, thus we
should still see a signal in the resulting model of the environmental variables that fulfill their resource needs.

Mean Akaike Information Criterion (AIC) values and the modified AIC for smaller sample sizes (AICc) were calculated for the 100 native range dataset runs and the 100 non-native range dataset validation runs, respectively. Within each range, runs were compared using best predictive accuracy as measured by the area under the receiver operating characteristic curve (AUC) values. The dataset for Ontario was selected based on the highest predictive accuracy values in an effort to increase the generalizability of the model; i.e. it was best able to distinguish between presences and absences, while perhaps sacrificing that it was not the ‘best’ model in terms of fit (i.e. lowest AIC values).

4.2.5.2. Generalized Linear Models
To determine the relationship between mute swan distribution and the landscape, I used generalized linear models. Variables were assessed for outliers, high leverage points, skewness, normality and heteroskedasticity. Some variables were log transformed to reduce their skewness (Smith, 2007). The logistic multiple regression models for a binomial dependent variable with a logit-link function were performed in R (R Development Core Team, 2012). Each model’s overdispersion was assessed and all models had acceptable overdispersion values near 1.

4.2.5.3. Model Selection
Akaike weights were determined for each model and used to weight these models for the predictions only if their weight was greater than 0.1 (Burnham and Anderson, 1998; Westphal et al., 2003).

4.2.5.4. Assessing Model Transferability
Model transferability can be assessed by assessing the AUC values and maximum Kappa values for the calibration range and for the reciprocal range (Randin et al., 2006). As used by Randin et al. (2006) to assess model transferability, I deemed a model to be successful if it had AUC values greater than 0.7 and maximum Kappa values greater than 0.4 in both ranges. I also
calculated the differences in probability values between the models applied internally and those applied externally, and spatially mapped these values (Randin et al., 2006).

4.2.5.5. Variable Importance

To determine the most important variables for each model, Nagelkerke’s $R^2$ values (and an adjusted version of Nagelkerke’s $R^2$ using the formula $1 – (n-1) / (n-p) \times (1 – D^2)$, where $n =$ sample size, $p =$ number of parameters and $D^2 = R^2$ (Guisan and Zimmermann, 2000)) were calculated for individual variables. These give a measure of the explanation of variance for each individual variable and can indicate the relative importance of each environmental variable.

4.3. Results

The means of the 100 runs of each model validated internally showed fair to good predictive accuracy based on AUC values between 0.69 and 0.81 (Araújo et al. 2005; originally used by Swets, 1988) (Table 4-2). Swets (1988) classified AUC values into classes including 0.6-0.7 as poor, 0.7-0.8 as fair, 0.8-0.9 as good and >0.9 as excellent. Overall, the LCCE models had higher predictive accuracy values when evaluated internally in Britain than did the LCC models. This did not hold true for the Ontario models.

<table>
<thead>
<tr>
<th>Theoretical Models</th>
<th>Britain Mean AUC</th>
<th>Ontario Mean AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCCE</td>
<td>0.81</td>
<td>0.69</td>
</tr>
<tr>
<td>LCC</td>
<td>0.79</td>
<td>0.75</td>
</tr>
<tr>
<td>Habitat + Climate</td>
<td>0.79</td>
<td>0.72</td>
</tr>
<tr>
<td>Land Cover</td>
<td>0.77</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Only the LCC and Habitat and Climate models built in Britain and Ontario were weighted for the final predictions as there were two models with Akaike weights greater than 0.1 (Tables 4-3 and 4-4).
Table 4-3 Evaluation of Land cover + Climate, Habitat + Climate and Land cover final models: Akaike weights.

<table>
<thead>
<tr>
<th>Theoretical Models</th>
<th>Britain Akaike Weights</th>
<th>Ontario Akaike Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCC</td>
<td>0.65</td>
<td>0.44</td>
</tr>
<tr>
<td>Habitat + Climate</td>
<td>0.35</td>
<td>0.56</td>
</tr>
<tr>
<td>Land Cover</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 4-4 Evaluation of Land Cover + Climate + Elevation, Habitat + Climate and Land cover final models: Akaike weights.

<table>
<thead>
<tr>
<th>Theoretical Models</th>
<th>Britain Akaike Weights</th>
<th>Ontario Akaike Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCCE</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Habitat + Climate</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Land Cover</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

With the exception of the Ontario LCCE model, all models evaluated internally had AUC values of greater than 0.7 and maximum Kappa values of greater than 0.4 (Table 4-5). These meet the requirements of Randin et al. (2006)’s internal evaluation for assessing model transferability. In terms of model transferability, the LCCE model built in Britain had a fair predictive accuracy in Ontario. Overall, the maximum Kappa values were poor for models evaluated externally, with the exception of the British weighted LCC and Habitat + Climate model which was the closest to meeting the requirements of maximum Kappa > 0.4. In terms of comparing models with and without elevation, the LCCE models applied externally both had lower predictive accuracy than the corresponding weighted LCC and Habitat + Climate predictions.
Table 4-5 Evaluation of model transferability between native and non-native ranges (comparing internal and external model validation measures such as AUC and maximum Kappa).

<table>
<thead>
<tr>
<th>Evaluation Range</th>
<th>Internal</th>
<th>External</th>
</tr>
</thead>
<tbody>
<tr>
<td>Models</td>
<td>AUC</td>
<td>Maximum Kappa</td>
</tr>
<tr>
<td>Native (Britain) LCC and Habitat + Climate models</td>
<td>0.82</td>
<td>0.49</td>
</tr>
<tr>
<td>Native (Britain) LCCE model</td>
<td>0.85</td>
<td>0.56</td>
</tr>
<tr>
<td>Non-Native (Ontario) LCC and Habitat + Climate models</td>
<td>0.90</td>
<td>0.74</td>
</tr>
<tr>
<td>Non-Native (Ontario) LCCE model</td>
<td>0.88</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Maps of model predictions of probabilities were made using only the models without elevation as the British weighted LCC and Habitat + Climate models had fair predictive accuracy values (Table 4-5), while the three other models performed poorly. Mute swan establishment in its native range of Britain has been mainly concentrated in the interior southeast part of the country with some establishment along the southern coasts and patchy establishment in the northern half of the country. Overall models calibrated in the non-native range under-predicted the probabilities of current mute swan distribution in its native range (Figure 4-1a and 4-1b). Curiously, models calibrated in the non-native range over-predicted probabilities of mute swan presence on the Northwest coast of Scotland and generally under-predicted probability of mute swan presence inland, with the exception of some urban areas. There was consistency across models in terms of predicting lower probabilities of mute swan presence in the far north of Britain, but there were large differences in their predictions of urban areas, especially in the southeastern part of Britain (the greatest density of mute swans) (see Chapter 3) (Figure 4-1c) where the non-native model under-predicted probability of mute swan presence, with the exception of more accurate predictions near London.
Figure 4-1 Predicted probability of mute swan presence in its native range of Britain using: (a) the LCC and Habitat + Climate model built in mute swan’s native range of Britain and (b) the LCC and Habitat + Climate model built in mute swan’s non-native range of Ontario. Differences in predicted probabilities determined by subtracting Ontario’s probability values from Britain’s probability values illustrated in (c).
Mute swan establishment in its non-native range has thus far been concentrated on the shores of the larger lakes in Ontario. The model calibrated in the non-native range demonstrates a strong correlation between mute swans and water bodies as well as urban areas. Grid cells which contain the shorelines of medium to large waterbodies have a higher probability of mute swan occupancy. Overall, the model calibrated in the native range had fair predictive accuracy and generally over-predicted the probabilities of current mute swan distribution in the non-native range (Figure 4-2a and 4-2b) though showed similar patterns in predicted probabilities. Differences in the probabilities of mute swan presence were mainly evident in major urban areas and on some peninsulas (Figure 4-2c). Both models show consistency in predicting a higher probability of mute swans in areas with higher percent water, on the coasts of the larger lakes and in urban areas, but the model calibrated in the native region under-predicted mute swan presence probabilities in some urban areas.
Figure 4-2 Predicted probability of mute swan presence in its non-native range of Ontario using: (a) the LCC and Habitat + Climate models built in mute swan’s non-native range and (b) the LCC and Habitat + Climate model built in mute swan’s native range. Differences in predicted probabilities determined by subtracting Britain’s probability values from Ontario’s probability values illustrated in (c).
Nagelkerke’s $R^2$ values of the individual environmental variables in Britain (native range) indicated that the percent urban land cover was the most important predictor variable, followed by elevation (Table 4-6). The most important environmental variables in Ontario (non-native range) were the mean annual temperature and percent urban land cover variables, followed by elevation. While the percent water and wetland variables were not the most important for either range, they did explain some of the loss of explained deviance for both ranges.

**Table 4-6** The percent loss of explained deviance when individual environmental variables are removed from the Land Cover + Climate + Elevation models of Britain (Nagelkerke’s $R^2$ value = 0.350) and Ontario (Nagelkerke’s $R^2$ value = 0.508).

<table>
<thead>
<tr>
<th>Environmental Variables</th>
<th>Percent Loss of Explained Deviance (%)</th>
<th>Percent Loss of Explained Deviance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Britain</td>
<td>Ontario</td>
</tr>
<tr>
<td>Percent Urban (%)</td>
<td>10.86</td>
<td>20.28</td>
</tr>
<tr>
<td>Percent Forest (%)</td>
<td>6.00</td>
<td>5.71</td>
</tr>
<tr>
<td>Percent Wetland (%)</td>
<td>0.00</td>
<td>14.57</td>
</tr>
<tr>
<td>Percent Water (%)</td>
<td>5.14</td>
<td>3.69</td>
</tr>
<tr>
<td>Mean Annual Temperature (°C*10)</td>
<td>3.71</td>
<td>37.60</td>
</tr>
<tr>
<td>Annual Precipitation (mm)</td>
<td>2.29</td>
<td>13.39</td>
</tr>
<tr>
<td>Elevation (m)</td>
<td>9.43</td>
<td>18.31</td>
</tr>
</tbody>
</table>

**4.4. Discussion**

A fundamental use of species distribution models is to predict species distributions either in another spatial or temporal dimension based on current species-landscape-climate relationships that are assumed to remain relatively constant. Modelling an invasive species’ distribution is more difficult than that of a native species as the species is likely expanding its range, and at a faster rate than it would be in its native range (Elith *et al.*, 2010). The issue of invasive species makes testing model transferability more complex, as it raises the question of whether we can hold the same standards for successful model transferability between native and non-native ranges in comparison to models transferred within regions of the native range (Graf *et al.*, 2006; Randin *et al.*, 2006). I evaluated these models for transferability by comparing predictive
accuracy and prediction patterns with current species presence probabilities; these were done reciprocally between ranges. I also assessed whether the same environmental variables were determined to be the most influential in both native and non-native ranges and particularly examined the role of elevation. While models do not meet the criteria for transferability in terms of predictive accuracy, there is a noticeable similarity in patterns of probability values when models are applied reciprocally (i.e. applying models built in native range to non-native range and vice versa), indicating that some similar relationships and habitat requirements are being found in both native and non-native ranges. While models built in the native range over-predict current mute swan distribution in the non-native range (based on probabilities), the patterns of prediction closely mimic that of the current distribution and the current model based on the non-native range.

At best, the models I tested exhibit asymmetrical transferability in that the model built in the native range is able to predict the non-native range with fair-good predictive accuracy while the model built in the non-native range shows poor predictive accuracy when applied to the native range. This is not surprising as Randin et al. (2006) found that the majority of 54 plant species in their study exhibited asymmetrical transferability within native ranges. I propose the following three explanations of asymmetrical transferability in my study: (1) differences in environmental variables between native and non-native ranges; (2) differences in biotic interactions between native and non-native ranges (such as population density pressures); and (3) inherent problems associated with predicting species that are undergoing range expansion.

While there are a broad range of values for the environmental variables of all of the data in the native range, there is less variation in the non-native range (results not shown). This may be a function of the smaller spatial extent of the non-native range, the smaller sample size of the non-native range or that mute swans have only established in areas within their non-native range that are the most ideal habitats, given their short history in Ontario. As a result of greater variability, models may predict higher probabilities of presence in the non-native range due to the areas in the non-native range meeting the requirements for mute swan presence. In contrast, models built based on non-native range data for a species that has not yet established in all possible locations is likely based on areas of introduction and higher habitat suitability, while areas with lower
habitat suitability are not accounted for. The results in this study are in contrast to those found by Fitzpatrick et al. (2007) who determined that models built in the non-native range over-predicted species distribution of fire ants in its native range as the fire ants had established in habitats where they had not established in their native range.

Species in all ranges exhibit local adaptation to their environments, while being constrained by their basic requirements. For example, mute swans may be found in habitats that may not provide all of their resources, but they are constrained to sites with freshwater (Wieloch et al., 2004). In their native range, mute swans are also constrained by their population density, which limits their access to ideal habitats and forces them to use less ideal habitats (Holm, 2002). For example, mute swans are often associated with agricultural fields in their native range which they have not yet been associated with in their non-native ranges (Kirby et al., 1994; Ciaranca et al., 1997). These habitat selection pressures in their native range again likely mean that they are found in less-ideal habitats which will affect the model.

There is an inherent difficulty in predicting a species that is expanding its range (Elith et al., 2010), and this difficulty is enhanced with an invasive species in a different environment than its native range which may not be occupying all suitable locations (Václavik and Meentemeyer, 2012). Randin et al. (2006) proposed a methodology to evaluating the transferability of a species but this has only been applied (as far as I know) to native species in their native habitats and does not address the issue of species with expanding ranges. It also depends on AUC and maximum Kappa, evaluative methods which have both been criticized (Liu et al., 2005; Lobo et al., 2007). However, while not fully transferable, models built in native ranges may be able to predict where areas of introduction or ideal habitat establishment will be in the non-native ranges (Broennimann et al., 2007; Fitzpatrick et al., 2007). Models built in the non-native range just may not have to the potential to be transferable to the native range.

Furthermore, there is no agreement on the best indictors of model accuracy. Probability values are a more accurate representation of model predictions than binary presence-absence outputs, especially when applied in novel ranges (Vaughan and Ormerod, 2005; Randin et al., 2006). Using AUC as a predictive accuracy assessment has recently come under scrutiny (Lobo et al.,
2007) as these values can be misleadingly high and the underlying predictive patterns can be quite spatially variable (Austin, 2007; Lobo et al., 2007). The use of a fixed threshold to determine presence-absences, though common in ecology, has been criticized for its inaccuracy.

Differences in the relative importance of certain important variables will also affect the models built, and hence the capacity for transferability between ranges. Calculation of Nagelkerke’s $R^2$ values revealed that urban land cover and elevation are important environmental variables in both ranges, indicating that these variables are consistently important in predicting mute swan distributions in both native and non-native ranges. The attraction to urban land cover in both native and non-native ranges is also exhibited by another invasive species, the house finch (Gammon and Maurer, 2002). The higher likelihood of mute swans being found near urban land cover in its native range is consistent with a number of studies in Britain that have concluded that mute swans are more often found in urban stretches than rural stretches of rivers (Mason et al., 2006), exhibit a high tolerance of pollution, hunting and fishing boats (Holm, 2002) and are found in smaller bodies of water characteristic of urban areas (Wieloch et al., 2004). Yet, it is highly probable that mute swans are not attracted to the urban areas themselves, but the associated waterfronts that fulfill their habitat requirements. In mute swan’s non-native range, urban areas experience urban heat island effects which temper their climate in the winter (Price, 1979) and may affect the warmth of water bodies within or near the urban boundaries, allowing them to freeze for a lesser period of time in the winter. In addition, mute swans that arrive in urban areas may be likely to stay since they do not need to locally migrate in the winter as they are in close proximity to suitable winter habitat with unfrozen water and food availability (Kirby et al., 1994). In mute swan’s native range, the importance of urban areas may also be a function of higher mute swan densities; possibly necessitating a need to find new habitats in novel areas.

While I would expect that aquatic (percent wetland and percent water) variables would be important environmental variables in both ranges as mute swans are an aquatic species, percent wetland was the only aquatic variable to result in a loss of greater than 10% of the explained deviance when removed from the model in the non-native range. Mute swans require water as they are an aquatic species mainly located in freshwater lakes, ponds and rivers and their main food source, SAV, is found in water (Kirby et al., 1994; Wieloch et al., 2004). While mute swans
have a relatively small territory size, and can abut adjacent mute swan territories with little conflict, a larger water body may provide greater food availability, have later freezing dates and possibly require less competition for territory space with growing populations (Wieloch et al., 2004). The lack of importance of percent water in the native range may be a statistical function or a function of scale, as Britain has much smaller waterbodies, mainly in the form of small lakes and rivers, whereas Ontario has much larger lakes and waterbodies. Therefore, the water variable measured at a 10 km scale in the native range may be a relatively small percentage of the grid cell, which is not the case in the non-native range of Ontario. While the aquatic variables did not explain the most lost in deviance in either of the two ranges, mute swan distribution predictions of both ranges reflected areas with water present.

The use of elevation as a environmental variable in species distribution models has been questioned (Austin, 2007), though has been used as a environmental variable in some studies of model transferability between native and non-native ranges (Mau-Crimmins et al., 2006; Fitzpatrick et al., 2007). The use of a environmental variable such as elevation, which may be very suitable as an environmental variable in the calibrating range, may not be suitable to be transferred to the non-native range, as a species’ relationship with elevation is more of a local function of its relationship with temperature and precipitation, and this actual local relationship between elevation and climate variables may not be the same when transferred to another range (Austin, 2007). To answer the question of whether the inclusion of elevation affects the accuracy of model prediction in the non-calibrating range, I compared the LCCE and LCC models. While LCCE models performed well when validated in their respective ranges, when models calibrated in the native range were used to predict the non-native range, their predictive accuracy decreased more relative to the LCC model. In addition, elevation was considered to be a more important variable in determining mute swan distribution in its native range than its non-native range. I thus recommend that models built for transferability purposes between native and non-native ranges do not include elevation as a predictor variable.

4.5. Conclusions
This study illustrates the challenges in transferring models between ranges, especially when modelling a species that is expanding its range in a potentially completely different environment.
Species distribution models that are applied reciprocally between native and non-native ranges are likely to exhibit asymmetrical transferability due to the different environments and relationships exhibited in different ranges, further complicated by adaptation and other biotic environmental variables.

My findings stress the importance of invasion establishment phase in building transferable models. While I conclude that these models are only asymmetrically transferable at best, they do exhibit some similarity in their prediction patterns when applied to the other range. In addition, the model built in the native range that over-predicts distribution in the non-native range but accurately predicts current presences suggests that species-landscape-climate relationships may remain consistent over space. When invasive species have not pervaded the landscape to the capacity that they could, models built in the native range can provide useful information as to where the most suitable habitats are within the non-native range. From a management perspective, the model predictions of mute swan distribution in the non-native range indicates that there are currently many more suitable habitats in which mute swans could establish and these locations should be prioritized in terms of conservation management. In particular, the native model as applied to the non-native range suggests that mute swans are likely to establish further inland along smaller waterbodies, and especially in urban areas.

I recommend further research of model transferability of invasive species including multiple species categorized by stage of invasion (or by a function of time since introduction). While model transferability can be determined by comparing predictive accuracy values between ranges, these rules may need to be modified for transferability of models of species in their native and non-native ranges in comparison to models of native species in different regions of their native range. Ideally, mechanistic models should be employed in tandem with species distribution models (Kearney and Porter, 2009; Kearney et al., 2010) as these may be able to provide further support to correlative species distribution model predictions.
Chapter 5
Impacts of Climate Change and Land Use Change on an Invasive Species’ Distribution

5.1. Introduction

To be able to effectively manage an invasive species, it is necessary to have an accurate assessment of its current distribution, an understanding of its relationship with the landscape and using these, a prediction of where it may establish in the future (Leung et al., 2006). The more that is understood about the current interactions between the landscape, climate and a species, the more accurately we can predict where a species will spread to in the future.

In the last decade, there has been an increasing popularity in applying species distribution models to predict future distributions of species under different scenarios (Wisz et al., 2008b; Wiens et al., 2009; Loiselle et al., 2010; Adams-Hosking et al., 2011). There is always uncertainty in predicting the future as we often must assume that the species-landscape-climate relationship stays relatively constant, cannot account for biotic interactions (unless incorporating a mechanistic model) and rely on climate change and land use change scenarios that have varying levels of uncertainty themselves (Elith et al., 2010). Many studies have examined the possibility of accounting for climate change while predicting future distributions of species (Araújo et al., 2005; Araújo and Luoto, 2007; Wiens et al., 2009; Loiselle et al., 2010; Adams-Hosking et al., 2011), while fewer have accounted for land use change scenarios (Wisz et al., 2008b; Loiselle et al., 2010). Validation of these scenarios when predicting forward is obviously problematic (Jeschke and Strayer, 2008); however, some have addressed this problem by back-casting to previous climates or land covers and species distributions to validate their models (Araújo et al., 2005). This approach relies on having full datasets for earlier time periods, which may not be available. While these predictions may have an inherent level of uncertainty, the goal is to determine, as accurately as possible, the location of future suitable habitats with the objective of managing the species and the landscape.
Habitat suitability maps can be produced by species distribution models built using presence-only data or presence-absence data, with newer modelling approaches reliably outperforming older approaches (Elith et al., 2006). Popular species distribution model approaches used to examine the relationship between species and the landscape include Generalized Linear Models, Maximum Entropy models such as MaxEnt, Classification and Regression Trees and Genetic Algorithm for Rule-set Production models (Godefroid and Koedam, 2007; Tsoar et al., 2007; Wisz et al., 2008b). These models are all based on the principles of correlations between environmental variables and species distributions (Kearney and Porter, 2009).

The advantages of correlative species distribution models are that they are versatile, can be applied to a variety of species data sets and they output statistical associations between species distributions and landscape environmental variables (Kearney and Porter, 2009). The disadvantages of these models are that they are dependent on the data used to create the model so may not be applicable to other, dissimilar landscapes, and that they do not examine mechanistic explanations for correlations, such as population dynamics or biotic interactions (Guisan et al., 2006; Kearney and Porter, 2009). In reality, the predictions of species distribution models are often limited by the variability of the training data, the accuracy of future scenarios, and species interactions and adaptation; environmental variables that are not included in the model (Buermann et al., 2008).

While MaxEnt has been often used in other fields (information technology for example) it has only recently been applied to species distribution predictions with great success (Dudik et al., 2007). MaxEnt is a machine-learning model based on a maximum entropy framework where, instead of using a regressive relationship, the model looks for a solution which is the most uniform between the species and the environmental variables (Phillips et al., 2006). MaxEnt therefore determines many possible distributions based on the environment-species correlations, and identifies the best potential distribution by selecting the distribution which is closest to uniform (maximum entropy) (Phillips et al., 2004). The advantages of this high-performance modelling method are that it relies on presence-only data, output is continuous, input data can be both categorical or continuous, it makes no prior assumptions on the distribution of the response curve and it is not as sensitive to sample size as most other algorithms (Phillips et al., 2006; Elith
et al., 2006; Buermann et al., 2008; Wisz et al., 2008a; Elith and Graham, 2009). The disadvantage of MaxEnt is that it has a tendency to over-fit data (although it does have a regularization function to prevent this) and predictions outside of the training data have high uncertainty (as do all predictive species distribution models) (Phillips et al., 2006; Wisz et al., 2008a).

In contrast, generalized linear models rely on a logistic relationship between the dependent and independent variables to determine a species’ relationship with the environmental variables and predict distributions. Generalized linear models are popular species distribution models and can be applied to presence-absence data (Randin et al., 2006; Broennimann et al., 2007; Zanini et al., 2009; Elith et al., 2010). When only presence data is available, MaxEnt has been found to have a better predictive performance than generalized linear models; however, when presence-absence data is available, it has been suggested that using a species distribution model designed for presence-absence data, such as a generalized linear model, is better (Elith and Graham, 2009).

Invasive species can have significant negative impacts on ecosystems in terms of displacing native species, reducing ecosystem integrity and reducing ecosystem resiliency to disturbances (Orians, 1986). Invasive species can also have significant impacts in terms of conservation and management budgets as more manpower and money will have to be allocated to eradicating the invasive species and managing the effects of the species on the ecosystem. The future of many species appears bleak with the realization that continued urbanization and climate change will have significant effects on the earth’s ecosystems by changing water cycles, habitat availability and inter-species dynamics, among other effects. However, invasive species are likely to thrive in these changing ecosystems due to their ability to capitalize on marginal habitats (Orians, 1986). Few predictions of invasive species distributions take into account land use change at a landscape scale using fine resolution predictions of urban expansion. There is a need to understand how the two major human-caused drivers of land cover change: urbanization and climate change, will affect the suitability of the landscape for invasive species’ habitats in order to more accurately predict future fundamental habitats. I address the following questions: (1) What is the increase in potential suitable habitat for the invasive mute swan by 2050 due to climate change and / or urbanization? and (2) Does climate change or land use change result in a
greater increase in sites with suitable habitat for mute swans by 2050? I examine these questions using the mute swan (*Cygnus olor*) population in Ontario as my case study species. Mute swans have recently expanded their range, have a rapidly growing population and are an invasive species in Ontario, therefore requiring a concerted management effort (Petrie and Francis, 2003).

5.2. Methods

5.2.1. Study Site
Ontario is a heterogeneous landscape with a mixture of land cover and land uses. In the south, the predominant land cover is agriculture and water with interspersed urban areas and forest. This gradient changes with increasing latitude, where forest and water become more prevalent. Ontario has undergone noticeable land use change within the last few decades, with an increase in urban areas, particularly in the Greater Toronto Area, and a loss of forest and wetlands in southern Ontario. For this study, I artificially limit the study extent to 49° north as my preliminary results conclude that mute swans, both currently and in the future, are not predicted to extend their range beyond approximately 49° north.

5.2.2. Mute Swan Distribution Data
Mute swans are an invasive species in southern Ontario where they have been established and breeding since approximately 1958 (Cadman *et al.*, 2007). They are considered to be invasive as they interact negatively with other waterfowl and consume and disturb a large amount of submerged aquatic vegetation, limiting its availability to other waterfowl (Petrie and Francis, 2003). The mute swan population has increased rapidly in the Great Lakes region of Ontario and the population is predicted to grow to be 16,000 by approximately 2050 (Petrie and Francis, 2003).

Mute swan distribution data were obtained from the 2001-2005 Ontario Breeding Bird Atlas which uses thousands of volunteers to collect bird breeding information from all of Ontario. The bird data used for this research were based on point counts (exact locations) of mute swans and rare bird data which were collected when volunteers are performing point counts or site visits and come across breeding evidence for a rare species (Bird Studies Canada *et al.*, 2006). Point
counts involve standing at a location for five minutes and recording any evidence of any birds (nesting, calling, or viewing) (Bird Studies Canada et al., 2006). This dataset gives a more accurate representation of a specific mute swan location than do presence-absence data aggregated over 10×10 km grid cells.

5.2.3. Sampling Effort

As MaxEnt chooses background points from the entire study extent unless otherwise specified, I accounted for this by using a bias grid for Ontario, such that MaxEnt would allocate more of the background points in areas that were more highly sampled and thus more likely reflect potential absence points. To do this, I interpolated the sampling effort provided by the Ontario Breeding Bird Atlas per 10×10 km grid cell. I should note that while point counts themselves had an equal sampling effort of 5 minutes per point count, their spatial distribution was skewed, absence data was unavailable and rare data did not conform to this sampling effort, and thus I determined that a bias grid was the best way to account for sampling effort.

5.2.4. Models: MaxEnt and Generalized Linear Models

I use the machine learning model MaxEnt to model current species-climate-landscape relationships and predict future distributions using climate change and urbanization scenarios. 10,000 background points were selected using the aforementioned sampling effort bias grid. I validated my models using 70% of the sample for training and a random validation sample of 30%. I employed bootstrapping for cross-validation and used 500 replicates per scenario (Adams-Hosking et al., 2011). I allowed clamping (a procedure that prevents the model from over-extrapolating beyond the training data’s range of environmental variables) and used a threshold rule of equal training sensitivity and specificity to determine presence-absence (Liu et al., 2005). I employed a jackknife procedure to determine variable importance. Finally, I stipulated a logistic output, meaning that these values are a predicted probability within the range of the training variables, or a relative suitability when applied outside the range of the training variables (Elith et al., 2011). I did all of the above twice, once using a regularization factor of 1 and once using a regularization factor of 4. Regularization factors control for over-fitting. The default value for MaxEnt is 1; however, this value is often considered to be too low (and restrictive) when predicting future climate or landscape scenarios in which the environmental
variable range for the predicted scenario exceeds that of the environmental variable range used to construct the model (Rodda et al., 2011).

I followed a similar procedure with my generalized linear models, in which I used 70% for training and 30% for validation. I created a pseudo-absence dataset, equal in number to the presence dataset (see Chapter 2) and using the same sampling effort bias grid used for determining background points in MaxEnt. A threshold rule of equal training sensitivity and specificity was again used to determine presence-absence (Liu et al., 2005). Variable importance was determined using Nagelkerke’s $R^2$ values (see Chapter 3).

5.2.5. Predictor Variables

Environmental variables that would best reflect possible landscape, climate and spatial environmental variables affecting habitat selection by mute swans were chosen. Current climate data, including annual mean temperature, mean temperature for the warmest quarter, mean temperature for the coolest quarter and annual precipitation averaged over the period from 1950-2000 data were obtained from the WorldClim dataset (Hijmans et al., 2005). These data have a resolution of 30 seconds (approximately 950 m). Spearman rank correlations indicated that the temperature variables were highly correlated. Preliminary analysis using MaxEnt indicated that the mean temperature for the coolest quarter variable had the greatest contribution to the model (however these relative contributions should be used cautiously when variables are highly correlated) and ecologically, this variable is most likely to be the most restrictive temperature variable for mute swan distribution and thus was retained.

I also utilized three different land cover maps to model current land cover in Ontario; I did this due to varying resolutions, extents and accuracies of the data sets. I first used Southern Ontario Land Resource Information System (SOLRIS) data (Ontario Ministry of Natural Resources, 2008) to model southern Ontario’s land cover as this is the highest resolution land cover data available for the province at 15 m. I then used Ontario Land Cover (OLC) data which is available at 25 m resolution for the portion of Ontario not covered by SOLRIS. OLC data is based on MODIS remote sensing (Ontario Ministry of Natural Resources, 2006). Unfortunately, interpretation of this data is not very accurate for urban land cover as bare soil is often
mistakenly classified as urban areas. Thus, I used GlobCover data, which is available at a 300 m resolution, to correct the accuracy of the OLC data in terms of urban land cover (European Space Agency et al., 2008). By reclassifying both the GlobCover and OLC data as a binary 0/1 for non-urban / urban, I was able to multiply the raster datasets together and deduce where urban land cover likely would be found in northern Ontario. The average percentage of this land cover per 950 m grid cell (the same as those used for the climate data) was then calculated. Percent water, percent urban and percent greenbelt areas and protected parks were determined and all other land cover (such as forest and agriculture) was designated as ‘other’ to be used as potential area for urban growth in the urbanization scenario.

5.2.6. Climate Change Scenarios

To predict climate change, I used ensemble A2 scenario climate predictions for the 2050’s time period, encompassing 2040 to 2069, based on the Intergovernmental Panel on Climate Change Fourth Assessment Report (Soloman et al., 2007). The A2 scenario reflects a future regionalized world with locally oriented development and a rapidly increasing population (Soloman et al., 2007). An ensemble forecasting approach was used whereby all 25 global circulation models were validated against the actual predictions for 1950-2000. Standard deviations of mean values of all 25 global circulation models’ annual mean temperature and annual precipitation predictions were calculated and used to determine the top five models whose models were most accurate in back-casting to the actual climate in 1950-2000. These five global circulation models were: Australia’s Commonwealth Scientific and Industrial Research Organization (CSIRO)’s MK3 model, the Canadian Global Circulation Model (CGCM3.1), the American’s National Aeronautics and Space Administration Goddard Institute for Space Studies (GISS) model, France’s Institut Pierre Simon Laplace IPSL model, and the German’s Meteorological Institute of the University of Bonn’s ECHO-G model. Un-weighted averages of these models’ predictions were then used as the future A2 scenario. These data, available at a 30 second resolution, were obtained from the International Center for Tropical Agriculture (CIAT), CGIAR Research Program on Climate Change, and Agriculture and Food Security (CCAFS) database for global circulation models and 1950-2000 averages were obtained from the WorldClim database (Hijmans et al., 2005; Ramirez and Jarvis, 2008).
5.2.7. Urbanization Scenarios

At the present time, no urbanization models are available for southern Ontario. Therefore, extrapolation from current greenbelt data, conservation data and population projections were used to approximate urban growth. This was mainly based on the growth plans for the Greater Toronto Area, which is the largest urban metropolis in my study area. The Greater Toronto Area has delineated a Greenbelt that envelopes the city and allows for an approximate doubling of urban land cover area over an unspecified amount of time (Ministry of Municipal Housing and Affairs, 2005). Population growth estimates for the 2050s suggest that there will be a doubling of current population size (population in year 2000 = approximately 11.7 million inhabitants while projected population in year 2058 = approximately 19.9 million inhabitants) (Ramlo et al., 2009). While it is difficult to reflect urban land use growth using population projections, as density of urbanization will be different depending on location, size of current city and need for high density housing in that city, I use population estimates and allowable urban growth area as an approximate estimation (which is an approximate doubling of urban area).

Urban growth predictions were created by determining the potential area of allowable urban growth per grid cell using the following formula:

\[
\text{Potential Area for Urban Growth} = \\
\text{Total Area per Cell} - \text{Water} - \text{Greenbelt} - \text{Protected Parks} - \text{Current Urban}
\]

I assumed that agricultural, forested and other lands were available to be converted to urban land uses, with the exception of waterbodies and wetlands, current parks and protected areas and the Greenbelt, which is a protected area of limited urban building surrounding the city of Toronto (Ministry of Municipal Housing and Affairs, 2005). Cells with a Total Available Area > 2 * Current Percent Urban allowed for an automatic doubling of urban land cover within each cell. Cells which had a Total Available Area < 2 * Current Percent Urban resulted in extra urban development which had to be allocated elsewhere. The allocation of extra urban land cover was mainly an issue in the Greater Toronto Area, due to its already high density of urban area and limited total available area for further urbanization. Thus, all extra urban land cover was
allocated to the areas surrounding metropolitan Toronto and Hamilton, within the Greenbelt. For the rest of Ontario, cells with urban land cover percentages of greater than 50% per cell were aggregated, to which 1 km buffers were applied to determine priority areas for extra urban growth allocation. Surrounding cells within the buffer were then allocated extra urban growth. Thus, urban growth scenarios were applied equally to each ~950 m cell, but then further growth for particularly dense urban areas tended to be applied primarily in larger urban cities instead of smaller urban towns.

5.2.8. Habitat Suitability Prediction Mapping

Future possible locations with suitable habitat for mute swans, accounting for climate change and land use changes by 2050 were mapped. Percentage area with habitat suitable for mute swans under each scenario was determined using the individually determined thresholds (see above) for presence-absence.

5.3. Results

*MaxEnt* models built using the training data from Ontario and validated using the same environmental variables (i.e. not predicting the future) had excellent predictive accuracy with AUC values > 0.9 (Swets, 1988). Validation of the generalized linear model’s training data yielded a good AUC value of 0.82 (Swets, 1988). Thus, all models were good models to use for future predictions with different ranges of environmental variables.

In *MaxEnt*, mean temperature of the coolest quarter (bio11) was determined to be the most important variable affecting mute swan distribution (Figure 5-1). This was followed by percent water (perwat) and distantly by percent urban land cover (perurb). When variables were individually removed, the highest loss of model fit was found when the temperature or water variables were removed. In contrast, there was less loss of the model’s predictive accuracy or power of explanation when the percent urban (perurb) or precipitation (bio12) variables were removed. The importance of each variable remained relatively consistent when performing a jackknife of the testing gain and of AUC (results not shown). Therefore, the temperature and percent water variables provided the most information for determining mute swan distribution.
when used individually in the model and were also the most important environmental variables in the model (Table 5-1).

![Figure 5-1](image.png)

**Figure 5-1** Importance of environmental variables for current mute swan distribution in Ontario as determined by a jackknife of the regularized training gain in *MaxEnt*.

**Table 5-1** Permutation importance of each environmental variable to the built model as determined by *MaxEnt*.

<table>
<thead>
<tr>
<th>Predictor Variable</th>
<th>Permutation Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Temperature of the Coolest Quarter</td>
<td>81.8</td>
</tr>
<tr>
<td>Annual Precipitation</td>
<td>0.8</td>
</tr>
<tr>
<td>Percent Urban Land Cover</td>
<td>0.5</td>
</tr>
<tr>
<td>Percent Water Cover</td>
<td>16.9</td>
</tr>
</tbody>
</table>

Variable response curves, in which the probability of presence is calculated relative to each individual variable’s values, varied greatly between variables. Mute swans exhibited a positive relationship with temperature whereby their probability of presence increased with average mean temperature of the coolest quarter until a possible threshold of approximately -3°C was reached (Figure 5-2a). Mute swans showed a high probability of presence in a narrow range of temperatures in this study site. Mute swans also exhibited a higher probability of presence at mid-range precipitation values (Figure 5-2b). Mute swans showed a high correlation with urban areas when the percent urban area ranged from approximately 10-40% of the land cover in a grid cell, but this slowly decreased as percent urban land cover increased (Figure 5-2c). Finally, mute swans exhibited a low probability of presence at lower percent water but this probability increased as percent water increased (Figure 5-2d). This positive relationship was exhibited until
approximately 90% water cover, as at this point, there would be little opportunity for territory establishment.
Figure 5-2 Variable response curves when each environmental variable is used individually in *MaxEnt* including (a) mean temperature of coolest quarter (bio11), (b) annual precipitation (bio12), (c) percent urban land cover (perurb), and (d) percent water cover (perwat).

The relative importance of the environmental variables as determined by *MaxEnt* was corroborated by the GLM that also determined that the mean temperature of the coolest quarter and percent water variables were the most important (Table 5-2).

Table 5-2 Percent loss of explained deviance in the built model when individual environmental variables are removed from the full GLM with a Nagelkerke’s $R^2$ value of 0.453.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Percent Loss of Explained Deviance when Variable Removed (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Temperature of the Coolest Quarter</td>
<td>49.44</td>
</tr>
<tr>
<td>Annual Precipitation</td>
<td>1.77</td>
</tr>
<tr>
<td>Percent Urban Land Cover</td>
<td>0.22</td>
</tr>
<tr>
<td>Percent Water Cover</td>
<td>7.51</td>
</tr>
</tbody>
</table>

Overall, the current number of sites with suitable habitat for the mute swan in Ontario was approximately 10.9% of the study extent as determined by *MaxEnt* using a regularization factor of 1 and 12.1% of the study extent as determined by *MaxEnt* using a regularization factor of 4.
Models of the current distribution generally illustrate higher probabilities of mute swan presence in areas which they are currently located, and generally predict higher probabilities of mute swan presence in areas near waterbodies and urban areas. The generalized linear model corroborated this general pattern but illustrated a higher probability of suitable habitats for mute swans throughout the southwestern region of Ontario and estimated the current number of sites with suitable habitat at 15.2% of the study extent (Figure 5-3c).
Figure 5-3 Maps of predicted suitable habitat for mute swans at present (2001-2005) using (a) *MaxEnt* with a regularization factor of 1, (b) *MaxEnt* with a regularization factor of 4, and (c) a GLM. Current mute swan presence locations used to train the models are also shown.
Overall, the inclusion of the climate change scenario in MaxEnt’s predictions resulted in an increase in the percentage of sites with suitable habitat being predicted in mid-Ontario. The A2 scenario of climate change resulted in a large increase in percent area of suitable habitat with approximately 33.1% of the study extent predicted to have suitable habitat for occupancy for the MaxEnt model using a regularization factor of 1 and 34.4% using a MaxEnt model using a regularization factor of 4 (Figure 5-4a; 5-4b). These results were supported by the GLM results, as inclusion of the climate change scenario also resulted in a large increase in the percentage of cells with highly suitable habitat for mute swans to 30.0% (Figure 5-4c). However, the predictions of suitable habitat did differ in the southwestern region of Ontario, where the GLM predicted some suitable habitat, as did MaxEnt with a regularization factor of 4 corroborated this, MaxEnt with a regularization factor of 1 predicted less suitable habitat.
**Figure 5-4** Maps of predicted suitable habitat for mute swans in 2050 using the A2 climate change scenario using (a) *MaxEnt* with a regularization factor of 1, (b) *MaxEnt* with a regularization factor of 4, and (c) a GLM.
When only urbanization scenarios were included, the MaxEnt (regularization factors of 1 and 4) and GLM predictions of areas of suitable habitat for the mute swan slightly decreased to 10.2% (regularization factor = 1) and 11.0% (regularization factor = 4) and slightly increased to 15.4% respectively of the current study extent (Figure 5-5a; 5-5b; 5-5c), which are percentages similar to their current predicted areas of suitable habitat.
Figure 5-5 Maps of predicted suitable habitat for mute swans in 2050 using the urbanization scenario using (a) MaxEnt with a regularization factor of 1, (b) MaxEnt with a regularization factor of 4, and (c) a GLM.
Model predictions accounting for both urbanization and climate change scenarios generally illustrated the same results and patterns as found for the climate change-only scenarios; however, the percentage of the study extent with predicted suitable habitat did increase for MaxEnt (regularization factor = 1) to approximately 37.0 % (Figure 5-6a) and for a regularization factor of 4 to approximately 37.8% (Figure 5-6b) and remained consistent for the GLM at approximately 30.1% (Figure 5-6c).
Figure 5-6 Maps of predicted suitable habitat for mute swans in 2050 using the A2 climate change and urbanization scenarios using (a) MaxEnt with a regularization factor of 1, (b) MaxEnt with a regularization factor of 4, and (c) a GLM.
MaxEnt also produces two maps which indicate areas of predictive uncertainty. This uncertainty occurs when future scenario environmental variables contain values which were not within the range of those used for training the model. In terms of climate, mean temperature of the coolest quarter values for future A2 climate scenarios in the southwestern and southern parts of Ontario were higher than those found in the training data (Figure 5-7a). Precipitation values were also predicted to be higher in some patches (Figure 5-7a). No variables were found to have values outside of the training range in terms of the future urbanization scenario (Figure 5-7b).

When values of environmental variables fall outside the range of those used in training the model, this creates uncertainty in the predictions as the model must extrapolate beyond its training data (this potential uncertainty is illustrated in purple on the map) (Figure 5-8a; 5-8b). Note the obvious correlation between the maps of temperature and precipitation values outside of the training range in Figure 5-7a and the map of highly uncertain predictions in Figure 5-8a.
Figure 5-7 Maps illustrating the limiting environmental variables for MaxEnt for the (a) climate scenario and (b) urbanization scenario.
Figure 5-8 Maps illustrating the levels of uncertainty in the predictions using the (a) climate scenario and (b) urbanization scenario.
5.4. Discussion
My study utilizes the invasive mute swan as a case study species to examine an invasive species’ relationship with environmental variables and predict how an invasive species’ distribution will change with increasing urbanization and climate change in its non-native range.

5.4.1 Response to Climate Change
Overall, climate change will result in increased temperatures in Ontario, thereby increasing the possible locations for establishment by the mute swan. Predictions of species distributions accounting for climate change vary between range shifts, range contractions and range expansions for different avian species (Cadman et al., 2007; Jarnevich and Stohlgren, 2009; Wiens et al., 2009; Adams-Hosking et al., 2011). In their study of all avian species in Ontario, Cadman et al. (2007) determined that species that are at the northern edge of their range in Ontario will be more likely to expand their range northwards with climate change. Both the generalized linear models and MaxEnt models agree that more suitable habitat for mute swans will be found further north by the 2050s.

All models indicate a range expansion, though this is more pronounced with the MaxEnt model using a regularization factor of 4 (again, decreasing over-fitting) and the generalized linear model’s predictions, while the MaxEnt model using a regularization factor of 1 indicates some shifts in distribution with areas of more suitable habitat being found northwards of the current mute swan presence points. This may be a function of the training data that does not include the full range of environmental variables of the mute swan’s range, and therefore does not account for the possibility that mute swans can be found at higher mean temperatures (as they are found throughout the northeastern United States). The values of these variables are outside the range of the training data for this study extent and thus may not be fully accounted for in model predictions. As the variable response curve shows a narrow band of temperature within which there is a high probability of mute swan presence, and as temperature is the most important predictor variable, which means it is the dominant variable in determining mute swan distribution (or suitable habitat) when predicting the future, it is likely that this narrow band of
predicted suitable habitat is a greater reflection of mute swan’s ideal habitat than potential habitat.

5.4.2 Response to Urbanization

Studies have shown varying responses of species to land use change (Wisz et al., 2008; Loiselle et al., 2010). As expected, those species that are currently associated with urban areas such as invasive species tend to show continued affinity to expanding urban areas (Múnoz and Real, 2006). Here, I find that mute swans are positively correlated to urban areas; however, the high probability of presence in urban areas decreases as percent urban areas increase. However, future urbanization scenarios do not result in a noticeable increase in the percentage of suitable habitat for mute swans, and in the MaxEnt models, incorporation of both climate change and urbanization scenarios resulted in a slightly lower percentage of suitable habitat than when only climate change scenarios were incorporated. Therefore, I believe that mute swans reach a threshold in their positive relationship with urban areas whereby availability of water and resources for territory establishment are diminished to the point where mute swans do not establish. Therefore, increasing the density of urban areas (i.e. percentage of urban areas per grid cell) may decrease the likelihood of mute swan presence.

The lack of response to urbanization may be a function of (1) the scale of study; or, more likely, (2) the lack of importance of the urban land cover variable in model. While we have found urban land cover to be an important predictor of mute swan distribution at multiple scales when buffers were centered around point data (see Chapter 2), the use of an imposed 1×1 km grid may somewhat remove the focus of urban areas as a function of grid cells being imposed randomly over points instead of as concentric buffers. While I believe that my urbanization scenario creation method accurately reflects both an increased densification and expansion of urban areas, this densification likely reduces the probability of mute swan presence in these areas, while expansion may increase probability of mute swan presence in others. Thus, the areas with suitable habitat for mute swans may only remain on the outskirts of urban areas, with loss of suitable habitat within urban areas, and is furthermore limited by the presence of usable water bodies. Furthermore, climate change affects all grid cells within the study extent, while the effect
of urbanization produces a localized response in urban areas only. Finally, while mute swans show a positive correlation with urban land cover, this variable was not found to be one of the most important or influential environmental variables in either the MaxEnt or GLM models. Therefore, while mute swans are often associated with urban land cover, this is not as influential as the climate or water variables in determining future suitable habitats.

5.4.3 Uncertainty in Predictions
The predictions made by MaxEnt illustrate uncertainty in the southwestern portion of the study extent where the mean annual temperatures are predicted to increase beyond that of the current mean annual temperatures. There is less uncertainty when only the urbanization scenarios are used as the changes in urbanization still remain within the range of the training environmental variable. Therefore the predictions based on urbanization scenarios are more dependable than those based on climate change scenarios. MaxEnt (with a regularization factor of 1) predicts little suitable habitat in these areas of high uncertainty, while MaxEnt (with a regularization factor of 4) and the GLM predict an increase in habitat suitability under all climate change scenarios. This is likely a function of the GLM calculating a relationship between environmental variables and mute swan presence that is less limited by the range of environmental variables than the maximum entropy method employed by MaxEnt (except when a higher regularization factor is used to reduce over-fitting). To address this uncertainty, future research with mute swans should endeavour to utilize presence data from the entire non-native range to train these models; however, at the current time there is no fine-scale point data available that has been collected with a consistent methodology for the entire non-native range.

5.4.4 Biotic Considerations
To knowledgably interpret these results, we need to account for biotic considerations (Araújo and Luoto, 2007), including the current spatial pattern of mute swan distributions in southwestern Ontario, range expansion at the northern edge of a species’ range, use of a limited climate niche to train the data, and the unknown effect of competition with other species. Accounting for the limited biotic information we currently have, it is likely that mute swans will expand their range north with changing climate as predicted by both models, but that they will
also continue to find suitable habitat within their current range. The relationship between species
and landscape (and climate) may be less definable at the northern edges of its range as they may
exhibit adaptation and plasticity which can alter species-landscape relationships, or at a
minimum, increase the variance in defining the species-climate-landscape relationship. As
MaxEnt and the GLM habitat suitability maps illustrate more desirable habitat north of the
current range, a range expansion northwards is likely. This will have implications for managers
who are not currently planning for mute swan arrival within their jurisdictions.

5.5. Conclusions
The GLM and MaxEnt models show consistency in their prediction of an increase in suitable
habitat north of mute swan’s current range in Ontario, indicating range expansion of the species
is likely. Although mute swans have not yet established in all currently available habitat in their
non-native range of Ontario, it is likely that their potential range will continue to expand as
global warming continues. Also, it is possible that mute swans will capitalize on a changing
climate by increasing their rate of range expansion and further work should study this potential.

Ontario’s urban areas are expected to steadily expand in coming decades and while this may not
increase the overall percentage of suitable habitat for mute swans, it will likely change habitat
suitability within current areas, with suitable habitat patterns following areas with waterbodies
and lower urban density. Thus, when considering future range expansion of mute swans in
Ontario, we must acknowledge that climate change will facilitate the survival and spread of this
invasive species. Invasion spread models will need to take these dynamic processes into account.
Additionally, as mute swans are highly correlated with urban land cover in their native range of
Britain (see Chapter 4), mute swan densities may change as a function of population dynamics
within urban areas, possibly requiring further management of urban water bodies and inter-
species dynamics.
Chapter 6
Summary and Synthesis

6.1. Summary

The overarching objective of this dissertation was to further examine the application of species distribution models to invasive species. Invasive species offer a unique challenge to species distribution models as they are often expanding their ranges and therefore, their relationship with the environment is dynamic, while species distribution models model the environment statically. In this dissertation, I tested the effects of modifying different input parameters into the species distribution model, including scale, sample extent, climate and urbanization scenarios, and range from which environmental variables were measured. I concluded that composite scale models and models constructed from extents with larger ranges of environmental variables resulted in more robust models, and that limited success can be achieved in transferring models between ranges.

This dissertation addressed six major objectives, which were to:

1. Determine whether single scale or composite scale models can best represent mute swan’s distribution;
2. Examine the effect of varying or limiting the range of environmental variables when building models for a native species;
3. Test the transferability of models between native and non-native ranges;
4. Compare predictions made by two different types of species distribution models;
5. Predict the relative impact of climate change and urbanization on an invasive species’ range;
6. Contribute to mute swan ecology knowledge and management efforts by specifically examining their relationship with the landscape and predicting areas of suitable habitat both at present and in the future.
The following sections summarize the main results of this dissertation. This summary is followed by a synthesis where I address how these results contribute to filling the knowledge gaps described at the beginning of this dissertation. I conclude with a discussion of future research directions.

6.1.1. The scale at which environmental variables best explain mute swan distribution

The issue of scale is pervasive throughout all biogeographical research. In terms of species distribution modelling, there is a specific issue of the type and scale of the dependent data (species distribution data) available. Depending on the type and resolution of this data, this may automatically dictate the scale at which one examines the independent variables used to explain the dependent variables. If the data are available at a high resolution or as point data, the independent variables can be measured at single or multiple scales as deemed appropriate.

Chapter 2 examined the use of single-scale and composite-scale (composed of environmental variables measured at different scales) models in determining mute swan distributions. Three scales of environmental variables were examined based on ecological information about mute swan’s spatial interactions with the landscape. Within the single-scale models, the global model including all environmental variables had the best fit and the best predictive accuracy at the 140m and 3000m scales. In terms of the global model being the best model, it is pertinent to note that this is not necessarily universally applicable as it is dependent on the number of variables used and their collinearity. I chose a reasonable number of variables to begin and rejected some based on correlation tests. Overall, I determined that composite-scale models had better fit and comparable, if not higher predictive accuracy than the best single-scale models, indicating that environmental variables act at different scales in determining mute swan distribution.
6.1.2. The impact of sample selection bias and spatial extent on model predictive accuracy

As with the resolution of the dependent variable available for species distribution modelling, the extent of the data available may not be controllable when independent data sets are used. This is often the case with bird surveys which are limited by ecological regions or by study boundaries set for a different purpose. The extent of this dataset and, more specifically, the range of each environmental variable it includes in addition to the inclusion of marginal habitat affect the models built.

Chapter 3 examined the effects of limiting and varying the range of environmental variables used to train a species distribution model. Regional extents were artificially limited by freshwater ecoregions and density of mute swans. The latter was used to reflect studies where one might only collect data from known species locations due to cost and effort restrictions. Models built using the full extent, or models that included more marginal habitat and whose individual environmental variables encompassed a wider range of values had better predictive accuracy and higher explanations of variance than models trained in sub-extents that included less marginal habitat and narrower ranges of environmental variable values. The aforementioned models were better equipped to predict species distributions under a wider range of environmental variable values when applied outside their training extent because the model was built with a wider range of environmental variable values. Thus, as expected, collecting species distribution data in areas of known occupancy and high density will not elicit the best models with the highest predictive accuracy.

6.1.3. Model transferability between native and non-native ranges

A key aspect of modelling invasive species distributions is the predictive capacity and transferability of models built in the native range and applied to the non-native range. This potential for transferability needs to be established for numerous species to determine whether there is a pattern, such that models for potential invasive species or very recently introduced species can be built in the native range and applied in the non-native range to determine areas at high risk of invasion. For a model to truly be considered as transferable, models built in the
native range and non-native ranges should both exhibit at least fair predictive accuracies when applied to the reciprocal ranges.

Chapter 4 establishes the potential for model transferability between both native and non-native ranges and the validity of including elevation as a predictor variable. The model built in mute swan’s native range of Britain and applied to the non-native range of Ontario exhibited fair predictive accuracy; however, they did over-predict mute swan’s distribution in its non-native range. These models exhibited asymmetrical transferability as the model built in the non-native range of Ontario and applied to the native range of Britain exhibited poor predictive accuracy and under-predicted mute swan distribution. In addition, while the inclusion of elevation increased the predictive accuracy and fit of the model when validated in its training range, the model built in the native range that included elevation had lower predictive accuracy when applied to the non-native range. Inclusion of elevation as an environmental variable may only be useful when the model is applied locally and when elevation’s relationship with temperature and precipitation is likely to be spatially consistent. However, when the model is being transferred between native and non-native ranges, it is unlikely that elevation’s relationship with temperature and precipitation will be consistent and inclusion of elevation as an environmental variable will result in a decrease in the model’s predictive accuracy.

Chapter 4 also illustrated that urban land cover was one of the most important environmental variables in determining mute swan distribution in both their native and non-native ranges, indicating that this species is also attracted to urban land cover in its native range, and that this is not a landscape relationship that occurs only in its non-native range. This has implications for our understanding of invasive species-urban areas relationships, which are commonly expected in non-native ranges but may occur also in native ranges. While Kirby et al. (1994) hypothesized that mute swans are more likely to be found in urban areas during the winter when other food sources are unavailable; it is especially interesting that this strong relationship exists in both ranges even for summer breeding habitats; this is presumably when food resources would be plentiful and easily attained in non-urban areas.
6.1.4. Using species distribution models to predict future distributions under climate change and land use change scenarios

Predictions of the future are inherently uncertain under any scenario as (approximate) validation is only possible through back-casting. True validation can only occur when we are able to measure actual species distribution as a function of its response to adaptation, dispersal, interaction with other species, land use change and climate change. Thus, predictions of future species distributions make the assumption that a species’ relationship with the landscape is unchanging temporally and use present relationships to predict future distributions. While climate change forecasts are abundant, there are few land use change projections available at a landscape scale. In chapter 5, I employed existing climate change projections, but created my own urbanization projections informed by current urban areas, projected population growth, and protected areas limiting growth, to predict future mute swan distributions.

Chapter 5 predicted future locations of suitable habitats for mute swans in their non-native range of Ontario. Using an ensemble forecast of the A2 scenario of climate change and an urbanization projection using an urban densification and sprawl formula, I created maps of potential suitable habitats in which mute swans could establish in the future, given that their relationship with the environment will not change. While there were some differences between model predictions, relative importance of each environmental variable in the models was consistent between MaxEnt and the GLM. Habitat suitability maps produced using both MaxEnt and GLM approaches indicate that climate change will have a more noticeable effect on potential mute swan range than urbanization. In addition, all models predicted that mute swans will expand their range northwards. Any indications of a northwards range shift are partly explained statistically, as the lower extent of the range of values for the climate variables of the training data for the southwest portion of the study extent exceeds that of the values of the climate variables predicted for the future. However, as mute swans are currently present south of the current study extent, it can be hypothesized that these temperatures will also be suitable for mute swan establishment in the future (as these temperature gradients generally shift northwards with climate change).
6.2. Synthesis

6.2.1. Consideration of scale in species distribution models

Before species distribution models are built, care must be taken to determine the implications of aggregation or summary of independent variables at whatever scale is chosen. Where possible, species distribution models should consider multiple scales and preferably composite scale models if these elicit the best fit and predictive accuracy. Unfortunately, species distribution models are generally limited by the resolution and type of species data. In addition, incorporating multiple scales into predictive species distribution models is difficult due to spatial autocorrelation issues when predicting species with summarized independent variables that may overlap. Regardless of whether the scale of independent variables can be chosen based on ecological knowledge, the impacts of the choice of scale on the resulting model and variable importance should be addressed.

6.2.2. Consideration of range of environmental variables in species distribution models

In terms of selecting a study extent for data collection, there is a danger in choosing study extents that are too closely sampled around known species distributions. If there are limitations to be placed on data collection due to cost or time, I suggest that sampling effort be more widely and less densely conducted to enable collection of data in both marginal and suitable habitats. This is very important in terms of transferability studies in which models are built in the native range and applied to the non-native range. While data are unlikely to be available from the entire native range, the models trained in the native range should be trained using data from both marginal and suitable habitats, and using a wide range of environmental conditions. Therefore, the variability and heterogeneity of the independent environmental variables measured, in addition to the inclusion of species presences especially in marginal habitats, will result in a model with better predictive accuracy and explanation of variance.
6.2.3. Consideration of transferability of species distribution models

Studies have shown that models transferred between regions within the native range have exhibited asymmetrical transferability (Randin et al., 2006), and therefore it is not surprising that models transferred between native and non-native ranges would exhibit asymmetrical transferability at best. In building models in both the native and non-native range and assessing for model transferability, the time since invasion and the stage of establishment of the invasive species in its non-native range should be considered. Invasive species that have been recently introduced and that likely have not dispersed to all available suitable habitats will not be in equilibrium with the landscape and thus models built in the non-native range will not take into account the possible variability in the environmental variables necessary for models with better predictive accuracy, as discussed in Chapter 3. As such, models built in the native range where the species is likely to be at equilibrium will predict areas of highly suitable habitat (or potentially areas of initial establishment or introduction) in the non-native range. Depending on the stage of the invasion, these models may over-predict the current distribution if this species has only recently established in the non-native range.

6.2.4. Consideration of predictions of species distributions when accounting for climate change and land use change

Ensemble forecasting using multiple species distribution models to predict species distributions in the future has recently become more popular as a method of reducing uncertainty in model predictions (Araújo and New, 2006). To reduce uncertainty in species distribution models’ predictions of the future, it is necessary to take into account the range of the environmental variables used to train the model, the validity of future scenario predictions and the parameters used to control the model predictions (for example, the inclusion of clamping procedures or regularization factors). Inclusion of two different regularization factors for the same model resulted in slightly different predictions of future habitat suitability for mute swans. Further consideration of the impact of regularization values of MaxEnt, and averaging of predictions made by species distribution models would potentially result in future predictions with less uncertainty.
Chapter 5 illustrated the complexity of the relationships found between mute swans, climate and land cover (particularly urban land cover). The models built using the *MaxEnt* approach indicated that mute swan distribution was positively correlated with urban land cover, though this was not the most important variable at this scale (in contrast with Chapter 2). In particular, Figure 2-c illustrates the positive correlation between mute swans and urban land cover that begins to slowly diminish at 40% percent urban land cover. Determination of this relationship was very important in explaining the lack of impact of urbanization on mute swan distribution, and the considerable impact of climate change (particularly increasing temperatures) on mute swan distribution in the future.

### 6.2.5. Contributions to mute swan ecology

I conclude that mute swans, an aquatic invasive species in their non-native range of Ontario, are positively correlated with water bodies, notably those found in urban areas. Also, when environmental variables explaining mute swan’s distribution are measured at a single scale, these should be ideally measured at the size of their territory. However, if composite models are feasible, environmental variables explaining mute swan distribution should be measured at the aforementioned ideal scales at which mute swans likely perceive these variables. While urbanization will have a localized effect on mute swan distribution within urban areas, climate change will have a much broader effect, likely resulting in a range expansion northwards. As mute swans are predicted to increase their population rapidly in the next few decades, and their range is predicted to expand based on habitat suitability predictions, conservation authorities must prepare strategies to deal with this invasive species, which will likely increase in density and spread northwards in Ontario.

### 6.3. Future Research Directions

While this dissertation has contributed to filling many of the knowledge gaps that exist in terms of modelling invasive species distributions, it has also revealed further knowledge gaps and research routes that could further contribute to improving invasive species distribution models. The following is a discussion of a few of these potential future research directions.
6.3.1. Species distribution models and connectivity

While species distribution models are extremely useful for creating risk maps and predicting future species distributions by taking into account environmental variables like climate change and urbanization, species distribution models do not determine the pathways by which range expansion or invasion may occur. Gravity models are becoming more commonly used in invasion ecology as a method of predicting the route of invasion of recently established species (Leung et al., 2006; Bossenbroek et al., 2001). A gravity model would provide a more accurate prediction of the pathways and dispersal from wetland to wetland of mute swans. These predictions are useful for management endeavours, especially in terms of devising short-term management priorities. In addition, some researchers have proposed a method of incorporating connectivity into species distribution models for habitat protection (Foltête et al., 2012). I would propose that this approach can also be applied to invasive species, especially in terms of mute swans, a species that has defined habitat patches (water bodies). This would add an extra dimension accounting for connectivity and potential movements between patches to predicting suitable habitats. By determining which landscape features are the most influential in the mute swan’s distribution, and by modelling connectivity and the potential pathways and habitats at risk of invasion by mute swans, managers will have more insight into how that species spreads throughout the landscape and what the key management priorities should be in limiting this spread.

6.3.2. Dispersal Rate

Species distribution models mainly focus on a species’ relationship with environmental variables; however, mechanistic processes can also be included in these models. In terms of future distribution predictions of mute swans, it would be useful to assess the potential validity of these predicted range expansions accounting for dispersal rate. This could be done by examining the distribution data collected by the 1981-1985 Ontario Breeding Bird Atlas (Cadman et al., 1987). To further verify the validity of my risk maps, one could map the 1981-1985 and 2001-2005 mute swan presence points in Ontario, measure the rate of spread and determine if mute
swans will be biologically capable of expanding their range as much as the risk maps predict in the next 40 years.

6.3.3. Multi-Species and Larger Spatial Extent Transferability Studies

Just as ensemble forecasting with multiple species distribution models or multiple climate change models can increase the accuracy of prediction, we can use multiple species to confirm certain hypotheses such that they can be universally applied with more confidence. For example, Chapter 4 determined that mute swans exhibit asymmetrical transferability when models are built and applied in their reciprocal native and non-native ranges. This conclusion of asymmetrical transferability would be more robust in application to invasive species if this theory could be applied to multiple species, and preferably to species for which we have data that extends over their entire native and non-native ranges.
References


DMTI CanMap Route Logistics (2011). DMTI Spatial CanMap Route Logistics DMTI Spatial Inc.: Markham, Ontario.


117
European Space Agency et al. (2008). Globcover Land Cover Product European Space Agency et al., ESA’s Earth Observation Data User Element (DUE), Europe.


