INCORPORATING PLANT COMMUNITY STRUCTURE IN SPECIES DISTRIBUTION MODELLING: A SPECIES CO-OCCURRENCE BASED COMPOSITE APPROACH

A Thesis Submitted to the College of Graduate and Postdoctoral Studies

In Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy

In the Department of Plant Sciences

University of Saskatchewan

Saskatoon

By

ANJIKA UDAYANGA ATTANAYAKE

Permission to Use

In presenting this thesis/dissertation in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis/dissertation in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis/dissertation work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis/dissertation or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis/dissertation.

Requests for permission to copy or to make other use of the material in this thesis in whole or part should be addressed to:

Head of the Department of Plant Sciences

College of Agriculture and Bioresources

University of Saskatchewan

51 Campus Drive

Saskatoon, Saskatchewan S7N 5A8, Canada

OR

Dean - College of Graduate and Postdoctoral Studies

University of Saskatchewan

116 Thorvaldson Building, 110 Science Place

Saskatoon, Saskatchewan S7N 5C9, Canada

Canada

Abstract

Species distribution models (SDM) with remotely sensed (RS) imagery are widely used in ecological studies and conservation planning and the performance is frequently limited by factors including small plant size, small numbers of observations, and scattered distribution patterns. The focus of my thesis was to develop and evaluate alternative SDM methodologies to deal with such challenges. I used a record of nine endemic species occurrences across 350km^2 from the Athabasca Sand Dunes in northern Saskatchewan to assess five different modelling algorithms including modern regression and machine learning techniques to understand how species distribution characteristics influence model prediction accuracies. All habitat modelling methods showed robust performance (>0.5 AUC), with the best performance in most cases from generalized linear models (GLM). I looked at how habitat predictions can be influenced by different threshold probabilities. That analysis highlights that actively selecting the optimum level is the best approach compared to the standard high threshold approach as with the latter there is a potential to deliver inconsistent predictions compared to observed patterns of occurrence frequency. I assessed the dune environment by evaluating dune morphologies, long-term dune spatio-temporal variations, and rates of woody vegetation encroachment and dune stabilization to evaluate an important potential threat to the Athabasca endemic flora. The Athabasca sand dunes are currently active and characterized morphologically by crescentic ridge and morphodynamically by transverse form dunes. The net extent of dune stabilization between 1985 and 2014 was 53.76 km^2 or nearly 20 percent of the total open sand dune extent. Continuing stabilization of the Athabasca sand dunes region may present conservation concerns for these narrowly distributed endemic taxa.

The development of composite-SDM framework used small-scale plant occurrence and UAV imagery from Kernen Prairie, a remnant Fescue prairie in Saskatoon, Saskatchewan. The effectiveness of the five algorithms was evaluated in light of the distributions of the species and the spatial resolutions of predictors. My testing clearly showed that each method was capable of handling a wide range of low to high-frequency species with strong GLM performance irrespective of the species distribution pattern. Critically, my work shows that, although GLM is computationally efficient, the method does not compromise accuracy for simplicity. I assessed plant community structure using image clustering methods and found that object-based clustered predictors and direct reflectance predictors followed very similar accuracy patterns indicating

limited advantages of using high-resolution images. The study found for high-frequency species (i.e. >0.5 or present in greater than 50% of plots) that prediction accuracy declines to be as low as the accuracy expected for low frequency species (i.e. <0.2 or present in less than 20% of plots). Higher prediction confidence was often observed with low-frequency species when the species occurred in a distinct habitat that was visually and spectrally distinct from the surroundings. This is in contrast to species widespread in different grassland habitats where distinct spectral signatures were lacking. The study has strong evidence to state that the optimal algorithmic performance is tied to a balanced number of presences and absences in the data. The co-occurrence analysis also revealed significant co-occurrence patterns are most common at moderate levels of species occurrence frequencies. The research does not indicate any consistent accuracy increase or decrease between baseline direct reflectance models and composite-SDM framework. Although accuracy changes were marginal with the composite-SDM framework, the method is well capable of influencing associated type 1 and type 2 error rates of the classification.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr Eric G. Lamb for his continuous encouragement and guidance throughout my time in the PhD programme. I would like to extend my appreciation to my advisory committee - Dr Yuguang Bai, Dr Christian Willenborg, Dr Xulin Guo and Dr Steve J. Shirtliffe for their criticism and guidance. I would be grateful to have Dr Joseph A. Veech as my external examiner, and his comments and critiques undoubtedly improved the quality of the thesis.

I would like to acknowledge the financial support provided by various institutions including the National Science and Engineering Research Council, the Government of Saskatchewan, Saskatchewan Ministry of Tourism, Parks, Culture, and Sports, and Saskatchewan Ministry of Environment to implement various field activities. Furthermore, I would like to express my sincere appreciation to all the scholarships I received including the Robert E Redmann Memorial Graduate Scholarship (2014-2017), the F.V. MacHardy Graduate Fellowship in Grassland Management (2014-2015), the Roderick Alan McLean Memorial Scholarship (2015), and the Devolved Postgraduate Scholarship of the Department of Plant Science (2015-2017).

I would like to acknowledge every person who was a part of field surveys carried out in the Athabasca sand dune region and the Kernen prairie. I cannot think of all names, but without them, the data collection is an impossible task to complete. They faced all weather conditions throughout the time in the field and extended their sincere motivations to move things forward without hesitation.

I would like to thank all of my friends and lab mated for their friendship over the last few years and their invaluable thoughts on various things positively impacted to guide my journey.

Finally, I would like to extend sincere gratitude to my parents and spouse for their invaluable support and encouragement to reach my life goals.

Table of Content

List of Tables

List of Figures

List of Abbreviations

Area Under the Curve	AUC
Artificial Neural Networks	ANN
Atmospheric Correction	ATCOR
Bi-Temporal Layer Stack technique	BTLS
Classification and Regression Trees	CART
Coefficient of Variation	CV
Dark Object Subtraction	DOS
Digital Numbers	DN
Enhanced Thematic Mapper+	ETM+
False Negative	FN
False Positive	FP
Generalized Additive Models	GAM
Generalized Linear Models	GLM
Geographic Information System	GIS
Histogram Minimum Method	HMM
Landsat - Operational Land Imager	OLI
Landsat - Thematic Mapper	TM
Landsat Ecosystem Disturbance Adaptive Processing System	LEDAPS
MODerate resolution atmospheric TRANsmission	MODTRAN
Multivariate Adaptive Regression Splines	MARS
Post-classification Comparison Change Detection	PCCD
Radiative Transfer Code	RTC
Random Forest	RF
Receiver Operating Characteristic	ROC
Remote Sensing	RS
Species Distribution Model	SDM
Terra - Advanced Spaceborne Thermal Emission and Reflection Radiometer	ASTER
Terra - Moderate-resolution Imaging Spectroradiometer	MODIS

Top-of-Atmosphere	TOA
True Negative	TN
True Positive	TP
True Skill Statistics	TSS
United States Geological Survey	USGS
Unmanned Aerial Vehicle	UAV

Chapter 1

1 General introduction

1.1 Overview

Understanding the spatial and temporal distribution patterns and trends of biological organisms is an important aspect of ecology and conservation management. In general, species of conservation concern present challenges as small numbers of locations scattered distribution patterns and differences in realized niche intensify inaccuracies in the prediction process (Santika 2011, Zurell et al. 2016). As a result, incomplete species distribution data with small sample sizes create statistical power issues that may reduce species distribution model (SDM) robustness (Stockwell and Peterson 2002, Thuiller et al. 2004, Guisan et al. 2006a, Pearson et al. 2007, Wisz et al. 2008). Recent SDM studies take an interdisciplinary approach combining geographical and ecological perspectives into a single framework that positively influences the accuracy of the final results (Guisan et al. 2006a, Gogol-Prokurat 2011, Cord et al. 2014a, Pollock et al. 2014, Rocchini et al. 2015, Thorson et al. 2015, Thuiller et al. 2015, Zurell et al. 2016). The main effort of my research is to combine community assembly processes and remote sensing (RS) techniques together to fit SDMs with acceptable accuracy for conservation and ecological mapping purposes.

Remotely sensed data contains biophysical measures useful to identify the spatial and temporal characteristics of plant habitat, detect vegetation phenology, and vegetation structural changes (McKee 1979, Paisley et al. 1991, Tsoar 2005, Neigh et al. 2008, Hesse 2009, Ewing and Kocurek 2010, Wood et al. 2012, Cole et al. 2014). The effectiveness of these remote sensing tools and approaches are well characterized, and widespread use has been encouraged by the low-cost availability of remotely sensed databases (U.S. Geological Survey 2018). RS data sources provide more distinct details of the plant community establishment (realized niche) in the landscape compared to climate and environment based estimates of the larger suitable habitats (Cord et al. 2014a, He et al. 2015). The development of SDMs with RS predictors are more likely to produce models that account for plant community diversity patterns (Cord et al. 2014a, Rocchini et al. 2015), mainly because remotely observed reflectance of a plant community is primarily a result of different plant species composition across the landscape.

Plant community assembly processes are complex because more than one mechanism is generally acting on species establishment in a given locality (Guisan and Zimmermann 2000, Miller 2010, Thuiller et al. 2014). Therefore, the fundamental goal of many ecological studies has been to understand how various bio-physical processes interact to generate observable patterns of community structure (Whittaker 1967, Brooker et al. 2008). It is possible to infer in many cases that changes in community structure indicate changes in processes influenced by particular biological organization patterns. The initial conceptualization of SDM and model characterization is based on the environmental niche of a species (Gauch and Whittaker 1981, Austin et al. 1990, Austin and Gaywood 1994, Guisan et al. 2002, Thuiller et al. 2004). The modelling process uses climatic and edaphic variables most likely to represent the preferred habitat or environmental niche of the species. This conceptualization does not take into account biotic interactions in the modelling framework. The usefulness of biotic associations have not been explored extensively in the SDM context; however, measures of such biological associations may provide indicators useful in the fitting of species distribution models (Ulrich and J. Gotelli 2007, Ulrich and Gotelli 2010, Zimmermann et al. 2010, Veech 2013, Thuiller et al. 2015, Tobler et al. 2019). Exploring the utility of such ecological information, specifically co-occurrence patterns, in SDM is the core goal of my thesis.

Species co-occurrence patterns and similar ecological structure measures are thought to be the result of deterministic biotic and abiotic processes and can improve traditional SDMs (Ulrich 2004, Ulrich et al. 2017). Although such community-level modelling is promising, only a few approaches to this have been tested in the recent past (Pollock et al. 2014, Cazelles et al. 2015, D'Amen et al. 2015, Strona and Veech 2015, Buckley et al. 2016, Tikhonov et al. 2017, Tobler et al. 2019). Species co-occurrence analysis based on community-level matrices are widely tested methods (eg. Gotelli (2000), Ulrich and Gotelli (2010), Ulrich et al. (2017)); however, these community-level values provide little information that can be used to improve an SDM. The method proposed by Veech (2013) to test for significant patterns of co-occurrences identifies non-random co-occurrences (either positive associations or segregations) between pairs of species (Veech 2013), and has generally proved reliable in follow-up testing (Lavender et al. 2019). Pairwise metrics are attractive for incorporation into SDM modelling as they provide straightforward probabilities of association that can readily identify the species that may provide relevant distribution information about a focal species.

1.2 Objectives and outline of the thesis

The overall objective of my thesis is to test the capabilities of direct SDM methodological procedures and composite-SDM approaches to handle low to high occurrence-rate data with RS predictors. I evaluated a hybrid approach to SDMs that accounts for plant community structure (pair-wise species co-occurrence) with observed ecologically meaningful RS predictors (i.e. object-based analysis versus pixel-based). Object-based analytical methodologies are developed to minimize image classification inaccuracies, providing emphasis on the spatial organization of pixels rather the spectral features of individual pixels. I am specifically evaluating the potential to incorporate pairwise species co-occurrence information into species distribution models to enhance the prediction accuracy for rare and common plant species. My thesis is presented with two data chapters. The first data chapter presents an evaluation of direct SDM methodological procedures and a target species habitat spatio-temporal dynamics assessment to elaborate the usefulness of RS tools. The second chapter presents the composite-SDM development process, an assessment of object-based derived predictors to account for plant community structure, and an assessment of predictor spatial scale influence on SDM performance.

In the course of my thesis, I test the following specific hypotheses. 1) That choice of SDM algorithm will influence the prediction accuracy of species, and that the effect will be variable across species with contrasting prevalence patterns. 2) That remotely sensed predictor variables (reflectance bands) will contain sufficient distinct details necessary to model the habitat difference and the distribution of plant species (common and rare) across a landscape. 3) That the prevalence pattern of a species (high versus low) will significantly impact SDM prediction accuracy irrespective of the modelling algorithm. 4) That a choice to use a higher probability of presence thresholds (above 0.7) to produce binary maps will positively influence the predictive performance of SDM models. 5) That object-based predictor will positively influence the accuracy of predictions compared to the pixel-based predictor. 6) That model predictive performance will improve along a gradient of predictor spatial scale from low to high resolution, particularly for smaller bodied plant species. 7) That plant species frequency of occurrence will not influence patterns of co-occurrence with other species. 8) That the plant co-occurrence pattern observed will indicate or be associated with variation in plant community structure and composition across a landscape, and therefore will positively contribute to improving predictive performance in a

composite-SDM framework. I evaluate the above hypotheses in two broad studies as described in the specific objectives outlined below.

The first specific objective of my thesis was to evaluate the performance of different species distribution modelling techniques across species with contrasting prevalence patterns. In Chapter 3, I evaluate this general objective by 1) modelling the distribution of nine endemic species found in the Athabasca Sand dunes of Northern Saskatchewan using LANDSAT image data as predictor variables and a range of different modelling algorithms. These species include 7 that are relatively common and widely dispersed within the Athabasca sand dunes region, and 2 rarer species that specialize on a narrow set of sub-habitats, 2) examining how habitat characteristics, limited occurrence data, and different modelling techniques affect predictions of habitat suitability, 3) determining the most suitable modelling techniques or combinations of methods to precisely identify occurrences of these rare plant species, and 4) estimate species distributions and population sizes for each of the species, and suggest biodiversity conservation strategies based on mapping of the species distributions.

The second specific objective of my thesis was to integrate pairwise species co-occurrence details and ecologically meaningful predictors into grassland species distribution modelling. In Chapter 4, I evaluate this general objective by 1) estimating the relative strengths of SDM predictions for common and rare grassland plant species using ecologically meaningful RS predictors (i.e. object-based analysis versus pixel-based), 2) assessing the influence of species distribution frequency on the prediction accuracies, 3) testing the utility of probabilistic co-occurrence analysis approaches to identify sets of co-occurring species suitable for a Composite-SDM framework, and 4) evaluating the effectiveness of composite-SDM performance with integrated co-occurrence analysis to optimize over and underestimation of the prediction extent for common and rare species respectively.

Chapter 2

2 Literature review

2.1 Overview

The plant community assembly process is complex and unpredictable as more than one mechanism is often acting on species establishment in a given locality and time (Whittaker et al. 1975, Thuiller et al. 2015). A fundamental goal of ecological research is to understand how these interactive mechanisms generate observable patterns of species within ecosystems. The community assembly approach identifies plausible links between the process and patterns whereby changes in interaction processes are identified that likely indicate changes in observed patterns within the ecosystem. Plant community differences thus represent predictable biophysical environment gradients (Zimmermann et al. 2007, Cord et al. 2014a, He et al. 2015). The effort of characterizing such distinct variation in observed species occupancy patterns in the ecosystem using various indicators is the fundamental basis for Species Distribution Modeling (SDM).

This literature review first provides the theoretical basis for species habitat occupancy modelling including the historical development of the niche concept, theoretical justification of species distributions using the niche, and an explanation of common and rare species distributions. The second section of the literature review covers the general background of species distribution modelling with subsections covering the meaning of species distribution modelling, what is being predicted, and what sort of data are required for modelling tasks. The third section of the literature review is devoted to the species distribution modelling process and algorithms. This section includes the steps involved in SDM and detailed discussion of statistical models behind both modern regression and machine learning techniques. The fourth literature review section covers model evaluation and field validation. This section includes reasons for evaluating predictive performance, threshold-dependent measures of accuracy and threshold-independent measures of accuracy. The last section of the literature review is devoted to remote sensing aspects of SDMs and includes data sources, characteristics of RS data, image processing, image interpretation, and pixel-based vs object-based pattern recognition.

2.2 Species habitat occupancy

2.2.1 Historic development of the species niche concept and Hutchison's n-dimensional hypervolume

Species distribution modelling draws fundamentally from an understanding of the concepts of the species niche and species-environment relationships. The niche concept is an important component of individual species behaviour, morphology and physiology. The concept is furthermore used to explain community-level species participation in ecosystem structure and function. Chase and Leibold (2003) reviewed the historical development of the niche concept including Darwin and Wallace's works on natural selection and evolution that they implicitly refer to species roles in the environment. There many early naturalists who have discussed the concepts related to a species niche, but Chase and Leibold (2003) view the primary founder of the niche concept to be Grinnell and his work on species abiotic ratios, habitat occupancy, food, and competitor relationships. Hutchinson (1957) formally defined the niche as the mathematical combination of environmental factors (the hypervolume) within which a species can survive and reproduce. Hutchinson further explained how to differentiate the fundamental (physiological or potential) niche and the realized (ecological, actual) niche in two classic papers (Hutchinson 1957, 1959).

The definition and the process of quantification of the niche concept by Hutchinson was a significant contribution to the literature. Hutchinson's original niche definition talks about a region of an n-dimensional hyperspace from the sum of all the environmental factors acting on the organism (Hutchinson 1957, 1959, Chase and Leibold 2003). According to this explanation, species $\boldsymbol{S_1}$ survival depends on two independent variables ($\boldsymbol{x_1}$ and $\boldsymbol{x_2}$). The variables defined are $\boldsymbol{x_1'}$ and $\boldsymbol{x_1''}$ for $\boldsymbol{x_1}$ and $\boldsymbol{x_2'}$ and $\boldsymbol{x_2''}$ for $\boldsymbol{x_2}$, considered limiting values permitting a species $\boldsymbol{S_1}$ to survive and reproduce. Thus, all possible environmental conditions permitting the species to exist was defined by the area of limiting coordinates. The shape of the area was defined based on the level of independence of each variable in consideration. The "n-dimensional hyperspace" or "n-dimensional hypervolume" was first formally described in Hutchinson (1957) classic paper. The hypervolume was defined as introducing variables $\boldsymbol{x_3}$ to $\boldsymbol{x_n}$ such that each variable represents ecological factors required for the species to exist (Hutchinson 1957). The limiting values of each

variable specify the hypervolume $\boldsymbol{N_1}$ which was defined as the fundamental niche of $\boldsymbol{S_1}$ (refer Figure 2.1). The realized niche space in multidimensional scale was defined as a volume $\Delta \boldsymbol{N}$ that changes over time, mainly because the values of some environmental factors are likely to change frequently (Hutchinson 1957).

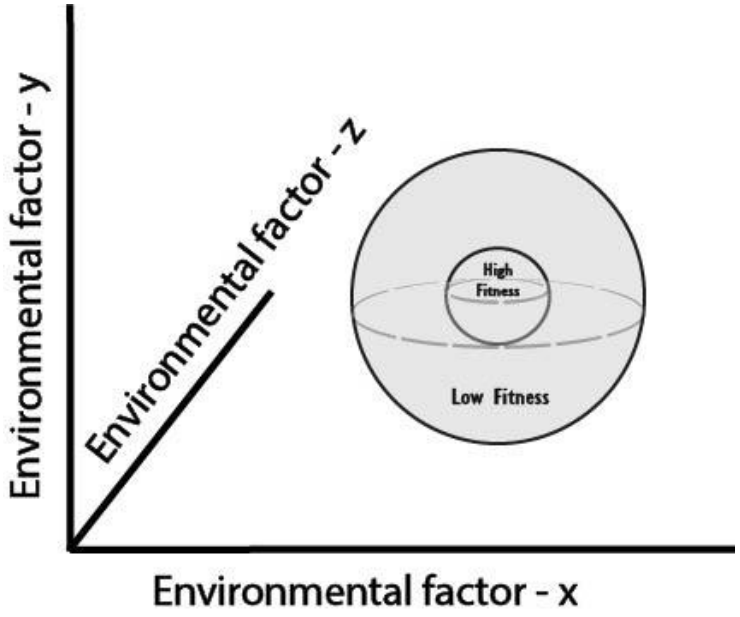


Figure 2.1: N-dimensional hypervolume as defined by Hutchison's (1957). The illustration is three-dimensional (three factors) scenario for simplification. The full length of each variable creates the volume that represents the total variation of each ecological gradient. The volume defined by the favorable conditions for a given species to survive is the fundamental niche.

2.2.2 Theoretical justification of species distributions and the niche

According to the historical advancement of the niche concept (Grinnell 1917), it is apparent that the several issues must be resolved in order for the concept to have a useful analytical role in ecology. For example, it is necessary to include other influencing factors for a species to exist in addition to resource competition. The synthesis by Chase and Leibold (2003), for example, defines the niche as "the joint description of the environmental conditions that allow a species to satisfy its minimum requirements so that the birth rate of a local population is equal to or greater than its death rate along with the set of per capita effects of that species on these environmental conditions". In addition to the above definition, Pulliam (2000) introduced an emphasis on positive

population growth rates as a measure identifying the species niche. Further, Pulliam (2000) suggests that a combination of the theoretical definition of niche "n-dimensional hypervolume" (Hutchinson 1957), metapopulation theory (Hanski 1998), and source-sink habitat theory (Pulliam 1988) together can elaborate a more meaningful relationship between niche of a species and distribution. Considering all the above theoretical justifications, the species niche refers to the biotic and abiotic environmental requirements that satisfy the existence of a species. Nonetheless, many species not only respond to variation in the abiotic environment, also they directly interact with one another, or more often indirectly interact by limiting resources (Pulliam 1988, Pulliam and Danielson 1991, Pulliam 2000).

Taking the ideas described above, Franklin and Miller (2009) graphically illustrated these concepts in geographical space (**G**) rather than environmental space (Figure 2.2). In this conception, **A** is the geographical area where the population growth rate is positive ($\boldsymbol{r_0(x) \geq 0}$). Identification of the niche boundaries in geographical space and/or the impact of species on their environment is made by lines of zero population growth rate ($\boldsymbol{r_0(x) = 0}$) (Chase and Leibold 2003, Holt 2009). This area is the fundamental niche (Pulliam 2000, Franklin and Miller 2009, Holt 2009) as defined by Hutchinson (1957). **B** is the area where a species either has a competitive advantage over other species or can coexist with those species. **M** is the area accessible to a species and potentially colonized via dispersal. $\boldsymbol{J_0}$ is the occupied "source" habitat and $\boldsymbol{J_p}$ is a suitable habitat that is unoccupied, potentially colonized if no dispersal limitations. Sink habitat is identified as places where a species is found in portions of M that do not intersect with both A and B together. Regions where population growth rate is negative ($\boldsymbol{r_0(x) < 0}$) are considered to be outside the niche (Holt 2009). Species observations (presences/absences) are the primary details going into the modelling framework. Thus, the realized niche represented by $\boldsymbol{J_0}$ **is** most likely the source of presences in the SDM context. Although, $\boldsymbol{J_0}$ provides details of presences, presence predictions might be in the $\boldsymbol{J_p}$ geographical area, predictions that would ultimately be classified as false positives.

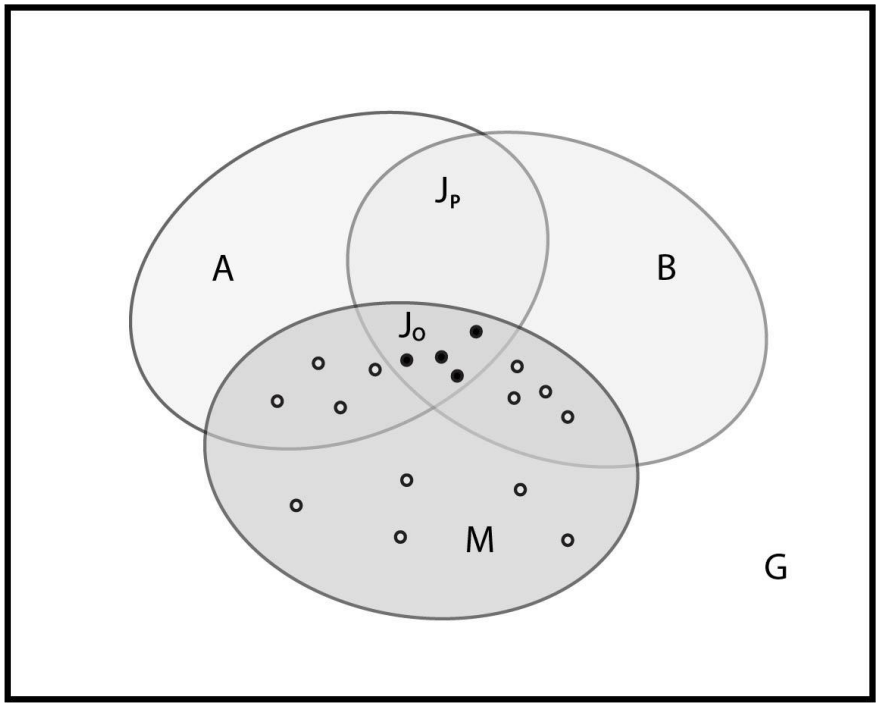


Figure 2.2: Representation of species niches in geographical space, re-drawn from Sober'on 2007 and Franklin and Miller (2009). G is representing a geographical space where species can exist, rather environmental gradients influencing the existence. The area A represents the geographical boundaries of positive population growth rates. The species can coexist with competitors within the boundaries of area B. The area labelled M is characterized by the species' dispersal ability within given timeframe. The occupied area is J_o and J_p is the potential occupied area if the species dispersed. Filled circles represent species existence in source habitat compared to sink habitats represented with open circles.

2.2.3 Species distribution variations – common and rare

The spatial and temporal distribution patterns of biological organisms are complex scenarios with many ecologists exploring the reasons behind observed variations (Grinnell 1917, Hutchinson 1957, 1959, Whittaker 1967, Whittaker et al. 1973, Gaston and Lawton 1990, Hanski 1998, Araújo and Guisan 2006, Guisan et al. 2006a, Hirzel et al. 2006, Thuiller et al. 2006, Elith and Graham 2009, Franklin et al. 2009, Miller 2010). In theory, the existence of a species can be explained by sufficiently favourable environmental conditions distributed along the geographical space, capacity to tolerate competition imposed by various external factors, and capability to spread through geographical space in order to maintain positive intrinsic population growth rate

(Hutchinson 1957, Kunin and Gaston 1996, Gaston et al. 2000, Pulliam 2000, Chase and Leibold 2003, Colwell and Rangel 2009, Franklin and Miller 2009, Holt 2009). Thus, a species can be common or rare in response to the above determinants. Nevertheless, species rarity or commonness is characterized by several important aspects such as the area of distribution, frequency of occurrence and size of the local population (Kruckeberg and Rabinowitz 1985, Soulé 1986, Kunin and Gaston 1996, Izco 1998). A commonly known factor is that ecological gradients must have acted over a long time to determine the distribution, frequency and abundance of a species currently observed.

2.2.3.1 Rare species distributions

Rare species are often of concern because many rare plants and animals are thought to be vulnerable to extinction (Kruckeberg and Rabinowitz 1985, Soulé 1986, Gaston and Lawton 1990, Kunin and Gaston 1996, Izco 1998, Gaston et al. 2000, Engler et al. 2004, Guisan et al. 2006a, Williams et al. 2009, Gogol-Prokurat 2011). There are multiple causes associated with rarity and endemism such as genetic factors, evolutionary history, and ecological influences (Soulé 1986, Kunin and Gaston 1996, Izco 1998). Common factors in rare species distributions are restricted geographic areas, specialized ecological requirements and isolation in specific habitats (Kunin and Gaston 1996, Izco 1998). In some instances, rarity is not a permanent scenario as a common species can become rare and endemics can become widespread with time (Kruckeberg and Rabinowitz 1985). Furthermore, observed rarity followed by commonness will occur at the beginning of every new species and not often recorded as it happens infrequently or is difficult to observe. The most common observation is that some rarities will remain rare in-definitely (Gaston 1996, Kunin and Gaston 1996, Gaston et al. 2000).

Kruckeberg and Rabinowitz (1985) organized the concept of rarity into categories based on local population size (large vs small), geographic range (large vs small), and habitat specificity (wide vs narrow). Dominant large populations of species can in some places be classified rare based if they are locally abundant over a large range in several habitats, locally abundant over a large range in specific habitats, locally abundant in several habitats but restricted geographically, and locally abundant in a specific habitat, but restricted geographically. Nondominant small populations of rare species can be classified into four categories based on constantly sparse over a large range and in several habitats, constantly sparse in a specific habitat but over a large range,

constantly sparse and geographically restricted in several habitats, and constantly sparse and geographically restricted in a specific habitat. More often the existence of a species is tightly bound to the environment that consists of biotic and abiotic factors. Usually, organisms do not occur where they cannot tolerate, but most often organisms do not occur in places where environmental conditions are favourable.

For example, a species can be specialized to maintain specific niche relationships to survive in predominant microclimate conditions. When local microclimate becomes highly discontinuous, some plants will be rare or endemic to the site. Microclimatic variations are highly associated with geological determinants that create discrete landforms (chemically and physically different substrates) provide better support for specialized species (Kruckeberg and Rabinowitz 1985). Another form of rarity is attributed with coevolution to adjust synergism or antagonism of interactive species (Soulé 1986, Kunin and Gaston 1996) and one of the coevolved pair of organisms might be rare (Ex: two symbionts, two commensals, herbivore-herb, pollinator-plant, host-pathogen, etc.). The process is considered a form of niche specialization to acquire necessary resources for survival and avoid competition to maintain positive population growth. Such niche specializations are usually a reason for a species to be rare (Kruckeberg and Rabinowitz 1985, Soulé 1986, Kunin and Gaston 1996, Gaston et al. 2000).

2.2.3.2 Common species distributions

Common species are those with high abundance as measured by criteria such as high local population size, large geographical range and wide habitat specificity (Kruckeberg and Rabinowitz 1985, Soulé 1986, Kunin and Gaston 1996). Common species generally exhibit large geographical coverage in several habitat types, implying that most common plant species have wider niche functions and an ability to tolerate large environmental variability, and often a competitive advantage that allows them to maintain positive population growth (Kruckeberg and Rabinowitz 1985, Kunin and Gaston 1996). If any species is competitively superior, they actively acquire necessary resources limiting the growth of inferior species in the same habitat. Generally, competitively superior species have high rates of habitat occupancy and ability to disperse to new habitats. Continuation of such expansions (for example in the case of an invasive species) can lead to competitively inferior species be rare or running to extinction. However, niche dynamics can happen completely opposite as stated by Kruckeberg and Rabinowitz (1985). For example,

changes in the environment may adversely influence the common species. This is an opportunity for the rare species in the same habitat to become wider spread as competition is reduced (Whittaker 1965, Kruckeberg and Rabinowitz 1985, Kunin and Gaston 1996, Izco 1998).

2.3 Species distribution models

2.3.1 What is species distribution modelling?

Species Distribution Modelling (SDM) is a major statistical method used to infer species distributions in space and time. SDMs have been used to describe either the niche of the species or to map suitable habitat for the species. The quantification of the species-environment relationships is the basic idea behind the prediction process. The initial stages of SDM development involve identifying species occurrences in the target ecosystem. This process requires extensive field surveys or extraction of data from already existing sources such as historic collection libraries or specimen collections (Guisan and Zimmermann 2000). Identification of biotic and abiotic factors that control species distributions are crucial for SDMs. In broad terms, climatic, topographic, and edaphic factors are significant contributors to determine species establishment patterns in the ecosystem. The SDM framework tries to establish plausible relationships between species occupancy patterns and environmental variables likely to influence habitat suitability. The model is a quantitative rule-based algorithm that finds the best fit between response and predictors. The model output is geospatial probabilities finally reflect species habitat preferences or environmental fit.

2.3.2 What is being predicted?

The primary result of a typical SDM is the prediction of the geo-spatial probability of occurrence of the target species. It is possible to predict biomass and species richness in addition to species presence-absence (Pineda and Lobo 2009, Dubuis et al. 2011, Zurell et al. 2016). There are various ways to interpret predictions of species presence and absence. For example, researchers have explained their models as predicting potentially occupied habitat (Guisan and Zimmermann 2000), habitat suitability (Hirzel and Guisan 2002, Hirzel and Arlettaz 2003), habitat use by the target species (Cassini 2011), species realized ecological niches (Guisan et al. 2006a, Thuiller et al. 2006, Soberón 2010, Anderson 2017), and species fundamental niches (Austin and Meyers

1996). Franklin and Miller (2009) argue that the prediction implies a species or habitat distribution if the model is based on presence and absence data. Interpolation and extrapolation are steps in any SDM; however careful consideration is a prior requirement based on expected temporal and spatial dynamics. Interpolation is defined as a process that fills in geographical information gaps where a species exists or is in equilibrium with the environment (Franklin and Miller 2009). Extrapolation is where predictions extend beyond the geographical sample space or into time periods beyond sampling timeline. This process provides predictions of suitable conditions for species to exist at various spatio-temporal scales (Franklin and Miller 2009).

2.3.3 Data for species distribution models

A requirement of all SDM techniques is the spatial coordinates of the observed occurrences of the target species and predictor variables characterizing those observations in environmental and geographic space (Franklin and Miller 2009). It is a necessity to have explicit information about where a species does not occur to distinguish appropriate and inappropriate habitat conditions. Such provision of absences to the modelling framework is also known as background data; in cases where hard (i.e. field verified) absences are not available pseudo-absence data should form the background (Wisz and Guisan 2009, Barbet-Massin et al. 2012). Pseudo absences are unverified absences of the target species generated using appropriate rules to provide locations where the species do not occur (Wisz and Guisan 2009). There are many studies that have evaluated the influence of sample size, generally finding that increasing sample size positively correlates with the performance of distribution models (Hirzel and Guisan 2002, Stockwell and Peterson 2002, Hernandez et al. 2006, Hjort and Marmion 2008, Wisz et al. 2008). The optimum sample size is an on-going debate (Elith et al. 2006, Wisz et al. 2008), and most likely depends on the specific context and objectives of the study. The prevalence pattern of the species is of great importance for SDMs, as abundance is often positively correlated with the sample size. The literature shows that the optimal balance between omission (false negatives) and commission (false positives) errors were achieved with a 1:1 ratio of presences and absences (McPherson et al. 2006). In some cases, the rarity of the target species makes for more predictable patterns if the rare species is a niche specialist highly associated with particular biotic and abiotic factors (Araújo and Guisan 2006, Franklin et al. 2009).

In general, while presence/absence modelling is most common, presence-only data modelling is a widely accepted avenue in SDMs. The biggest disadvantage with presence-only modelling is that the method can only predict the relative likelihood of species present at a site as the dataset lacks information about species prevalence. This leads one to conclude that presence/absence data produce more accurate predictions in comparison to presence-only data. However, it has been shown that a large number of randomly selected pseudo-absences, equally weighted to the presences, yielded the most accurate and reliable distribution models with generalised linear models, generalised additive models, multiple adaptive regression splines and classification techniques such as boosted regression trees, classification trees and random forests (Wisz and Guisan 2009, Lobo and Tognelli 2011, Stokland et al. 2011, Barbet-Massin et al. 2012). In addition, some literature suggests that the random selection of geographically and environmentally stratified pseudo-absences would have the greatest impact on the model's predictive accuracy when using classification and machine-learning techniques (Stokland et al. 2011, Barbet-Massin et al. 2012).

2.4 Modelling process and algorithms

2.4.1 Species modelling process

The process of SDM starts with observing species presences, identifying absent locations, and specifying associated predictor variables likely to influence or indicate the habitat occupancy. Then, the model training links predictors with species occurrences in geographical space to predict the likelihood of the species to be present or absent in un-sampled locations. The initial step in SDM is to develop a conceptual model that links the abiotic and biotic environment controlling spatial and temporal existence of a species. The process should thoroughly examine the biology of the target species to determine what variables to use as predictors of the presence of the species. There are various options available such as edaphic, topographic and climatic variables in broad terms. However, the final set of predictors should be determined based on the criteria of attaining the most certain predictive performance at a reasonable cost. Then, the data on species presences in geographical space and environmental factors representing species distributions are linked using a quantitative or rule-based model. Final steps involve applying the trained model to produce a map of predicted occurrences of a species and methods to evaluate uncertainties in the predictions.

Statistical modelling techniques, mathematical advancements and machine learning methods have all been applied to overcome the various challenges associated with SDMs, though a single best approach has not emerged yet. Scattered habitat occupancy patterns of the species and limited understanding of factors that result in specific distribution patterns are common challenges that contribute significantly to increase uncertainties associated with predictions (Williams et al., 2009). The conceptual modelling stage can involve various thought processes such as empirical/phenomenological modelling, mechanistic (process) models, or analytical (theoretical) models likely complementary to each other to deal with challenging species characteristics (Guisan and Zimmermann 2000). Algorithmically, statistical models – modern regression and machine learning methods are widely used methods for linking response-predictor relationships.

SDMs are a family of computational techniques that use a variety of different mathematical algorithms suitable for explaining the complex distribution patterns of species in the ecosystem. Regression techniques have been in use for decades and are the foundation of SDM methods. The cornerstone of modern regression is the generalized linear models (GLM) which is regularly used in situations where the distribution of the response variable is non-normal, but the explanatory variable-response variable relationship is linear (Guisan et al. 2002, Austin et al. 2006, Hirzel et al. 2006, Elith and Graham 2009, Dubuis et al. 2013). Generalized additive models (GAM) have been widely used in SDM given the flexibility those methods provide to identify and describe non-linear relationships between predictors and response (Yee and Mitchell 1991, Guisan et al. 2002, Yee and Mackenzie 2002, Guisan et al. 2006b, Williams et al. 2009, Rodríguez-Rey et al. 2013). A recent development in SDM is to use multivariate adaptive regression splines (MARS) which is a generalization of stepwise linear regression, well suited to problems with large numbers of predictor variables, or as a modification of the regression tree approach (Leathwick et al. 2006, Mateo et al. 2010).

Machine learning methods are also widely used in SDM; these are supervised learning mechanisms that construct functions directly from the training data using various kinds of algorithms (Breiman 1996, 2001). Decision trees (DT), which include classification and regression trees are a primary machine learning technique that has been used to seek patterns in data (Friedman and Roosen 1995, Breiman 2001, Elith et al. 2008, Olden et al. 2008). Artificial neural

networks (ANN) are similar and can hierarchically partition data into subareas focusing only on informative portions of the data. This can facilitate model fit in high-dimensional problems where it is very difficult to fit parametric response functions (Olden et al. 2008, Franklin and Miller 2009, Williams et al. 2009, Crisci et al. 2012). Maximum entropy (ME) is a general-purpose machine learning technique composed of statistical mechanics and information theory. ME states the best approximation of an unknown distribution as an observed probability distribution with maximum entropy, which is always subject to known constraints. Recent developments of the above method have shown high predictive accuracy compared to other SDM techniques where "presence-only" data was used in the modelling process (Elith et al. 2006, Phillips et al. 2006, Elith et al. 2011, Phillips 2012, Merow et al. 2013). Each of the major SDM techniques is discussed in detail in the next section.

2.4.2 Modern regression techniques

2.4.2.1 Generalized linear models (GLM)

GLM was considered the cornerstone as it lays the base for many other sophisticated modelling techniques (Austin, 2002, 2007; Brotons, Thuiller, Araújo, & Hirzel, 2004; Jane Elith & Leathwick, 2009; Antoine Guisan & Zimmermann, 2000; Miller, 2010). The method is often used as a baseline for comparison of prediction accuracy with other modelling techniques. This approach to modelling has given wider flexibility in handling various types of scales of measures such as binary, categorical, ordered categorical, ordinal, interval and ratio variables under a single theoretical and computational background. This is an excellent advantage in ecological data handling and in particular for SDM as the variables that characterize species distributions can come under any measurement scales. However, GLMs have shown inadequate performance in situations where plant species response curves for an environmental gradient deviate from approximate normal curves in the sense of being symmetric and bell-shaped.

2.4.2.2 Generalized additive models (GAM)

Introduction of generalized additive modelling (GAM) was based on the data-driven approach rather a priori assumed model-driven approaches to model species distributions in ecology (Yee and Mitchell 1991). This greatly facilitates determining the shape of the species

response curves and detecting features such as bimodality or pronounced asymmetry in the data. An important feature of GAM is that the modelling is not limited to linear relationships and additional flexibility for modelling non-linear relationships is acquired, replacing linear functions of the variables by unspecified smoothing functions (Hastie and Tibshirani 1990). In practical terms, smoothing function will be estimated from the data using techniques for smoothing scatter plots such as running-mean and running-line smoothers, running medians, regression splines, cubic smoothing splines, locally-weighted running-line smoothers (Hastie and Tibshirani 1990). Also, GAMs can handle various data distribution patterns such as binomial, Poisson, Multinomial, Gaussian, providing similar flexibility to the GLM in modelling species distributions.

2.4.2.3 Multivariate adaptive regression splines (MARS)

The multivariate adaptive regression splines (MARS) is a modification of the regression tree approach well suited to large numbers of predictor variables (high dimensional data). Friedman and Roosen (1995) state that the MARS procedure has the right strength to model relationships that are closely additive or involve interactions. The ability to integrate higher-order interaction terms is a great advantage over GAMs which are additive and present difficulties in introducing interaction terms (Franklin and Miller 2009). Methodologically MARS is well suited for presence-only data, however the method is flexible handling both continuous and categorical data (Leathwick et al. 2006, Barbet-Massin et al. 2012). MARS is capable of handling fairly big data files with minimum preparation work. It is not necessary to eliminate outliers in the data as recursive partitioning process separate data into disjoint categories and the effect of outliers are confined (Friedman and Roosen 1995). Such flexibility with MARS is advantageous compared to other regression techniques.

2.4.3 Machine learning algorithms

2.4.3.1 Classification and regression trees (CART)

The classification and regression trees (CART) approach is classified under tree-based methods, more specifically decision trees (DT) which are used in situations where the response is a categorical variable with more than two categories and where the predictors include both categorical and continuous variables (Franklin and Miller 2009). CART techniques have the great

advantage of accommodating missing values, avoiding the need for prior data transformations and elimination of outliers, ease of modelling nonlinear relationships, and inclusion of interaction effects between predictors (Elith et al. 2008). Fundamentally, tree-based methods produce a single best model which is considered to be a drawback (Elith et al. 2006). In the study, I will use a variant on CART, the boosted regression tree (BRT), a technique of boosting where large numbers of relatively simple tree models are combined iteratively to optimize predictive performance (Elith et al. 2008). The general criticism of CART methods is the lack of stability caused by high sensitivity to changes in observations (sample size) or varying the set of predictor variables. These sorts of changes finally lead to large differences in the predictions of the fitted models. The BRT capable of eliminating shortcomings associated with CART by boosting iterative stage-wise procedure to minimize lack of fit.

2.4.3.2 Artificial neural networks (ANN)

Artificial neural networks (ANN) are a collection of modelling techniques flexible enough to handle complex ecological phenomena with multiple interacting elements (Olden et al. 2008). Usually, GLMs and GAMs performances are profoundly affected by nonlinear high-dimensional data with non-additive circumstances (Franklin and Miller 2009). I used single hidden layer back-propagation (or single-layer perceptron) networks methodology in the study as it was commonly used ANNs in ecological SDMs (Thuiller 2003, Olden et al. 2008, Williams et al. 2009). The available literature suggests that ANNs achieve much higher classification accuracy than other available methods in complex classification problems. Nevertheless, it is possible to examine the contributions (both magnitude and direction) of the predictor variables compared to just regression coefficients from other methods (Olden et al. 2008).

2.5 Model evaluation and field validation

2.5.1 Why do we evaluate predictive performance?

Usually, any SDM analysis will result in at least some prediction errors. Such errors are associated with incomplete knowledge of the environmental factors influencing species distributions, lack of species spatial information, and measurement inaccuracies (Franklin and Miller 2009). Furthermore, specification or misspecification of model components could be

another source of errors, often these are associated with vague or ambiguous meanings of concepts, identification of correct predictor variables and estimations of parameters (Franklin and Miller 2009). Therefore, any forecasting model should be validated prior to implementing species predictions. In SDM practice it is not recommended to use the same data to both calibrate and evaluate the SDM, as overestimation of predictive performance is most likely to occur. New or independent data should be used to validate model performance. However, collecting a new set of data for model validation is often not feasible; the alternative would, therefore, be to partition the data into two portions. One set will be used to calibrate the model, called the training data, and another one will be used to validate the predictions, called the testing data (Guisan and Zimmermann 2000, Miller and Franklin 2002). In certain situations, studies implement bootstrap sampling (i.e. sampling with replacement) to generate a testing data set. This allows researchers to use all available data to estimate the final model parameters (Fielding and Bell 1997). Accuracy evaluation methods are either threshold-dependent and threshold-independent.

2.5.2 Threshold-dependent measures of accuracy

The final result of most modelling techniques is to produce a geographical probability surface that is then converted into pre-defined categorical classes using a threshold (Franklin and Miller 2009). Thus, the species presence and absence values are a conversion of the continuous geographical probability surface into a categorical one. The method is using a threshold and the result is a prediction of an occurrence of a species or a non-occurrence. Accuracy measures that depend on categorical predictions are called “threshold dependent measures” and are very often applied in SDM evaluations. Fundamentally, the concept is based on observed and predicted positives (presences) and negatives (absences) arranged as a two-by-two contingency table called “error matrix” or “confusion matrix” (refer table 2.1), similar in many ways to the familiar type I and II error charts used to consider null hypothesis statistical testing. There is a number of threshold dependent accuracy measures derived using the confusion matrix (refer table 2.2).

Table 2.1: An error matrix or confusion matrix for a two-class (binary) situation (Species presence versus absence). All values are counts based on the classification and the table adapted from Franklin and Miller (2009).

		Observed		
		Present	Absent	Sum
Predicted	Present	True Positive (TP)	False Positive (FP)	Total predicted presence
	Absent	False Negative (FN)	True Negative (TN)	Total predicted absent
	Sum	Total observed present	Total observed absent	Total number of observations (n)

Table 2.2: Formulae for threshold-dependent accuracy measures for species presence versus absence models based on the error matrix. Acronyms follow Table 2.1 and all formulas adapted from Franklin and Miller (2009).

Measure	Calculation
Sensitivity	TP / (TP + FN)
False negative rate	1 - Sensitivity
Specificity	TN / (TN + FP)
False positive rate	1 - Specificity
Percent correct classification (PCC)	(TP + TN) / n
Positive predictive power	TP / (TP + FP)
Odds ratio	(TP x TN) / (FP x FN)
Kappa	$\frac{[(TP+TN)-(((TP+FN)(TP+FP)+(FP+TN)(FN+TN))/n)]}{[n-((TP+FN)(TP+FP)+(FP+TN)(FN+TN))/n]}$
True skill statistic (TSS)	1 – maximum (Sensitivity + Specificity)

An evaluation of the associated costs of false negatives (omission) and false positives (commission) errors are central for SDM evaluation. Usually, there are various limitations associated with each statistical procedure mentioned. For example, the kappa statistic is a measure of categorical agreement that describes the difference between the observed agreement and chance agreement in theory, however, some literature strongly suggests that kappa might be sensitive to the prevalence patterns of species (Manel et al. 2001, Allouche et al. 2006, Freeman and Moisen 2008). The selection of the threshold totally depends on the choice of the researcher based on how the SDM will be used and the trade-off between false positives and false negatives. As an example, the threshold is set at 0.5 to produce binary maps for SDM models using logistic regression as a prediction algorithm in many different statistical software packages. Selection of the threshold balancing equal proportions of the false positive and false negative is the main consideration that leads to balance the prevalence of events in the sample (Fielding and Bell 1997, Manel et al. 2001, Freeman and Moisen 2008).

2.5.3 Threshold-independent measures of accuracy

Comparison of prediction accuracy among different modelling techniques is a practical situation in SDM research. Such situations are common when modelling rare plant species and common plant species together to reach better predictions. In such comparisons, it is recommended to use measures that are independent of the numbers of presences and absences (prevalence) in the sample, commonly called as threshold independent measures of accuracy (Franklin and Miller 2009). One of widely used threshold independent measure is the area under the curve" (AUC) of the receiver-operating characteristic (ROC) plot. The plot is a graph of the false-positive error rate (the x-axis) versus the true positive rate (the y-axis) corresponding to each possible value of threshold probability. Then, the AUC is calculated by summing the area under the ROC curve where the value can range from 0-1. If the value is greater than 0.5 means performance better than random. There are some literature suggesting that AUC is a reliable measure for model comparisons as AUC is not affected by changes in species prevalence (Manel et al. 2001). In addition to AUC, the pearson correlation coefficient has been used as a threshold-independent measure of predictive performance in SDM. The correlation measure is the degree to which predictions vary linearly with test data (Elith and Graham 2009, Phillips et al. 2009).

2.6 Remote sensing of environment

2.6.1 Popular applications and data sources

Species distribution modelling predictors are diverse and mainly depend on the data available to the researcher. The inclusion of remotely sensed (RS) predictors is a pragmatic approach in species distribution modelling considering the practical difficulties to measure and map more ecologically relevant predictors (Cord et al. 2013, Cord et al. 2014a). In the recent past, the use of remotely sensed data for various ecologically important purposes has been becoming popular, for example for the classification of land surface cover, detection of vegetation phenology, measurement of plant stress, surface soil moisture assessment, land and ocean surface temperature monitoring, and ocean vegetation stress assessment among other fields of applications (Hugenholtz 2005b, a, Ewing and Kocurek 2010, Hugenholtz et al. 2012, Mohamed and Verstraeten 2012, Wood et al. 2012, Cord et al. 2014b). One of the main reasons for the expansion of remote sensing applications is the no- or low-cost availability of remotely sensed databases (U.S. Geological Survey 2018). These databases include the Landsat - Thematic Mapper (TM), Landsat - Enhanced Thematic Mapper+ (ETM+), Landsat - Operational Land Imager (OLI), Terra - Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), Terra - Moderate-resolution Imaging Spectroradiometer (MODIS), and Aqua - Moderate-resolution Imaging Spectroradiometer (MODIS) image archives are available free of charge from USGS website (U.S. Geological Survey 2018). Among those archives, Landsat TM and ETM+ sensor images have been among the most commonly used to study vegetation dynamics such as effect of urban expansion on vegetation, characterization of vegetation biophysical properties, climatic influences, severe drought and subsequent recovery, irrigated agriculture expansion, insect outbreaks followed by logging and subsequent regeneration, forest fires with subsequent regeneration, sand dune migration analysis, understanding mineralogical variations in dune environments and determining dune morphologies etc. (Smith et al. 1990, Parker Gay Jr 1999, Levin et al. 2004, Howari et al. 2007, Yao et al. 2007, Özyavuz 2010, Markogianni et al. 2012, Mohamed and Verstraeten 2012, Wood et al. 2012, Dashtekian 2013).

There have also been increases in the use of remote sensing approaches for vegetation mapping and various other vegetation assessments (Norwine and Greegor 1983, Roughgarden et

al. 1991, Treitz and Howarth 1999, Markogianni et al. 2012, Wood et al. 2012). This is a very important application as land cover and land use changes strongly influence terrestrial biogeophysical and biogeochemical processes (Neigh et al. 2008). The theoretical basis for vegetation mapping is built on spectral profiles that show clear variation in the peak reflectance of the near-infrared wavelengths because of the internal structural organization of "green" leaves, and the highest absorption is in the red wavelengths because of the presence of chlorophyll (Jensen 2005). Various vegetation indices have been developed to remotely sense temporal and spatial variation in vegetation density, biomass, net primary production, plant community structure, stress level, decease severity, and pest impacts. The Normalized Difference Vegetation Index (NDVI) is considered a principal tool for measuring vegetation characteristics (Holben 1986, Neigh et al. 2008, Markogianni et al. 2012, Dashtekian 2013). Most of these remote sensing tools and approaches are well examined and have been demonstrated to be accurate at least under certain conditions.

2.6.2 Characteristics of remotely sensed data

Digital images are constructed based upon the energy reflected from objects in front of the sensor. A similar process is employed in remote sensing to measure the energy emitted from the earth's surface. Usually, the energy emitted is captured by a sensor mounted on spacecraft, aircraft, or unmanned aerial vehicle platform. The energy captured by sensors are mainly from either, reflected sunlight, radiating energy from the earth itself (passive remote sensing), or artificial energy sources such as a laser or radar (active remote sensing). The direct measure of emittance is called radiance, measured in watts per steradian per square meter (Jensen 2005, Campbell and Wynne 2011). In comparison, the reflectance is the ratio between the amount of energy leaving the target to the amount of energy striking the target (Richards and Jia 1999, Jensen 2005). The measure has no units and is a property of material being observed if all of the light leaving the target is intercepted by the sensor (specifically called hemispherical reflectance). In general, the measured radiance of an object depends on the illumination which is a function of the intensity of the energy source, the direction, the path of the energy source through the atmosphere, position of the object, and orientation towards the energy source. This measurement in remote sensing is called apparent reflectance, which is commonly used as reflectance. Usually, the sensor mounted

on the aerial vehicle measures spatial variations of the reflectance and record the data into discrete elements called pixels.

Sensors are specialized to capture electromagnetic variation in specific regions of the electromagnetic spectrum (bands). Quantized measures of reflectance are then converted into pixel elements that have discrete values in units of either reflectance or digital numbers (Jensen 2005, Chander et al. 2009). The image has distinct characteristics specified as spatial resolution, spectral resolution, radiometric resolution and temporal resolution (Richards and Jia 1999, Jensen 2005). Spatial resolution is a measure of the resolving power of the sensor to discriminate features in the field of view and is measured in pixel size. The detector size, focal length, and sensor altitude are key determinants of the area represented in a pixel. The spectral resolution is defined as the electromagnetic region that the sensor is sensitive. Multispectral or hyperspectral resolution is defined in terms of two wavelengths that specifies the location in the electromagnetic spectrum of the spectral band. Each band produces separate images with pixels corresponding to reflectance from specific electromagnetic regions. Radiometric resolution is defined as how sensitive the sensor is to brightness variations. The resolution is measured in bits and 8-bit depth can record 256 levels of brightness variation. Temporal resolution is the revisit time of the same area. The feasibility to use a remotely sensed image for a particular application are mainly determined by the four resolutions specifying imagery features.

2.6.3 Image preprocessing and interpretation

Remotely sensed images have a variety of information embedded in forms such as bands and various textures. Specialized knowledge is required to transform images into information; the process is called image interpretation. Usually, the interpretation process is not an easy task due to unusual perspectives – aerial view, spectral bands that go beyond the visible region of the electromagnetic spectrum, and unusual subject scales caused by differences in sensor resolution (Richards and Jia 1999, Jensen 2005). Therefore, image interpretation involves several steps: classification, enumeration, measurement, and delineation to overcome potential inaccuracies associated with the challenges mentioned above (Jensen 2005, Blaschke et al. 2008). The classification process is used to assign classes to imagery. This process involves feature detection, recognition, and identification to properly assign accurate classes. The enumeration task involves identifying and listing all distinct classes in the image. Quantification of a subject's distances and

height are one aspect of the measurement process. Moreover, quantitative assessment of the image pixel brightness is an integral part of the measurement. The distinction of subjects and clustered boundaries in the image are based on analysis techniques known as delineation. Combination of all of the above processes is required for proper image interpretation and often the process is challenged by gradual reflectance changes between clustered subjects. There are various aspects of the image that have been used to interpret subjects of interest. Those are image tone, image texture, shadows of subjects, patterns of recurring subject arrangements, feature associations, shapes of clustered subjects, the relative and absolute size of the target, and topographic position of the site (Richards and Jia 1999).

Often, remotely sensed images need to be preprocessed for solar, atmospheric, topographic, and sensor distortions prior to the principal analysis. The radiometric corrections, atmospheric corrections and geometric correction processes are of the utmost importance to implement accurate image-based analysis (Jensen 2005, Chander et al. 2008, Chander et al. 2009, Young et al. 2017). Digital images contain pixels scaled to Digital Numbers (DNs) for easy use. It is necessary to convert DNs to at-sensor radiance with brightness values in the physical units (Watts per square meter per micrometer per steradian) for subsequent comparative analysis. Sensor specific calibration coefficients for the conversion are available in the metadata comes with each image. Commonly available image processing software packages are capable of DNs to at-sensor conversion together with solar correction. The conversion of at-sensor radiance to top-of-atmosphere (TOA) reflectance is called solar correction (Jensen 2005, Young et al. 2017). Calculation of TOA reflectance usually accounts for exo-atmospheric solar irradiance, the distance between earth and sun, and angle of the sun. The measure is a ratio of reflected radiation from objects on earth to the incident irradiance upon the object of interest.

Atmospheric adjustments are used to minimize the influence to image brightness values from atmospheric distortions and sensor anomalies. Atmospheric scattering is mainly caused by particles in the atmosphere, specifically particle size and abundances are key variables to consider. However, scattering is strongly associated with the wavelength as well (Jensen 2005, Chander et al. 2009). Methodological procedures available to atmospheric corrections; radiative transfer code (RTC) analytical models, histogram minimum method (HMM) or the dark object subtraction (DOS) technique, image-based atmospheric effect minimization technique, MODTRAN

(MODerate resolution atmospheric TRANsmission), ATCOR, and Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) are some commonly used examples in the industry (Jensen 2005, Young et al. 2017). The geometric correction process involves positioning imagery to its right geographic location. Geometric rectification, image-to-image registration, and image-to-map registration are commonly used methods to reach greater geometric accuracy (Jensen 2005, Campbell and Wynne 2011, Young et al. 2017). The process of orthorectification is used to overcome the effect of relief and directional view of the sensor towards the object (Young et al. 2017).

2.6.4 Pixel-based image analysis and object-based pattern recognition

Image analysis involves describing image content in a way that can be used to draw meaningful judgements of objects (Jensen 2005). The image pixel is the basic analytical unit containing spectral details of objects. The historic development of pixel-based image processing is based on the concept of multidimensional feature space (Toutin 2004, Jensen 2005). If the analysis only uses pixels for generating details, the process obviously ignores the spatial context of the pixel. This is the main constraint identified in pixel-based analysis in comparison to the object-based or object-oriented image analysis (Blaschke 2010, Weng 2011, Blaschke et al. 2014). Whenever pixels are larger (coarse-scale imagery) or object size is close to the pixel size, its recommended to use per-pixel or sub-pixel image analysis (Blaschke et al. 2008). This scenario minimizes the spectral variability within each class and spectral mixing of each pixel, both together reduces classification inaccuracies. The biggest constraint with higher resolution imagery is higher within class classification variability due to the fact that one object is composed of many pixels (Blaschke et al. 2008, Weng 2011). There are many recent research and literature reviews highlighting classification inaccuracies caused by within object spectral variabilities associated with higher densities of pixels. This reasoning provides compelling evidence to move towards object-based methodological advancements for higher resolution image analysis.

Object-based analytical methodologies are developed to minimize high-resolution image classification in-accuracies, providing great emphasis on the spatial organizations of pixels rather spectral features of individual pixels. Generally, it is not possible to expect unique spectral properties of individual pixels, if the object in an image is composed of many pixels. However, spectral properties together with neighbourhood relations are likely to be unique for a given

scenario. The uniqueness of a neighbourhood is a result of variable colours, tones, and shadows that generate various patterns, shapes and textures(Jensen 2005, Blaschke et al. 2008, Weng 2011). Most algorithmic developments for object-based analysis are based on image segmentation principles generally categorized into point-based, edge-based, region-based, and combined methodologies (Blaschke 2010, Weng 2011, Blaschke et al. 2014). Generally, segments of an image are generated using one or more feature space dimensions such as mean, median, minimum, maximum, variance, coefficient of variation, etc. (Weng 2011, Blaschke et al. 2014). Incorporating such features with spatial dimensions usually minimize reflectance variation within objects and increase object-based feature extraction accuracies (Blaschke et al. 2014).

Image texture is a widely used object segmentation method with high and very high-resolution images (Blaschke et al. 2008, Weng 2011). The texture of an image or particular area is defined as repeating spatial arrangements of radiation intensities (Haralick et al. 1973, Ryherd and Woodcock 1996). Usually, the coarse texture is characterized as similar or higher within-class variability compared to between-class variability (Haralick et al. 1973). The smooth texture is observed where within-class variability is lower than between class variability (Haralick et al. 1973). This characteristic of the texture is a very useful feature to demarcate various plant community organizations in relation to community structure (Laliberte and Rango 2009, Wood et al. 2012). The moving window of user-specified spatial distances are frequently used to implement various texture calculation algorithms. Users can choose various forms of such algorithms, for example, range, variance, coefficient of variation, grey level co-occurrence matrix, maximum probability, contrast, homogeneity, and correlations are few well-recognized methods widely used to analyze image texture (Coburn and Roberts 2004, Laliberte and Rango 2009, Wood et al. 2012). The pixel-based moving window texture analysis usually creates difficulties to identify boundaries between different textures of an image if the objective is to segment various texture classes (Laliberte and Rango 2009). The problem is a result of boundary overlap between different texture classes and the problem intensify with increasing moving window size (Laliberte and Rango 2009). This is the biggest constraint for segmentation tasks using pixel-based methods in comparison to second-order statistics generated from grey level co-occurrence matrices. However, moving window operations can readily be used to enhance texture characteristics of an image that corresponds to various organizations of different objects in the field of view (Ryherd and Woodcock 1996, Laliberte and Rango 2009).

Chapter 3

3 Long-term sand dune spatio-temporal dynamics and endemic plant habitat extent in the Athabasca sand dunes of northern saskatchewan[1]

The manuscript entitled 'Long-term sand dune spatio-temporal dynamics and endemic plant habitat extent in the Athabasca sand dunes of northern Saskatchewan' (Chapter 3) was published in Remote Sensing in Ecology and Conservation. This manuscript was co-authored with Dandan Xu, Xulin Guo, and Eric G. Lamb. Dandan Xu and Xulin Guo provided critical methodological support for the remote sensing analyses in the study. Eric G. Lamb provided raw field data and supervised and guided all stages of project development, analysis and manuscript preparation.

[1] Attanayake, A. U., D. Xu, X. Guo, and E. G. Lamb. 2018. Long-term sand dune spatio-temporal dynamics and endemic plant habitat extent in the Athabasca sand dunes of northern Saskatchewan. Remote Sensing in Ecology and Conservation **5**:70-86.

3.1 Abstract

The Athabasca sand dunes in northern Saskatchewan and northeast Alberta are a unique landscape of moving sand that hosts nine narrowly distributed endemic vascular plant taxa. We modelled the extent of habitat for each species, corresponding dune morphologies in species habitat, spatial and temporal variation in dune environments, and rates of woody vegetation encroachment at dune boundaries to support an assessment of long-term threats for the Athabasca endemic dune flora. Landsat images were used to maximize the time spans and areal coverage of the study. The Athabasca sand dunes are currently active and characterized morphologically by crescentic ridge and morphodynamically by transverse form dunes. Longitudinal sand movement parallel to the dune axis resulted in the creation of new dune areas along the east and southeast boundaries of the dune fields at a rate of 0.14 km^2 $year^{-1}$. Forest succession along the western boundaries of the dune fields resulted at an annual dune loss of (1.98 km^2 $year^{-1}$). The net extent of dune stabilization between 1985 and 2014 was 53.76 km^2 or nearly 20 percent of the total open sand dune extent. All habitat modelling methods showed robust performance (>0.5 AUC), with the best performance in most cases from generalized linear models. The estimated total available/occupied habitat was comparatively low for the least abundant species *Achillea millefolium* (38.92 km^2) and *Armeria maritima* (48.82 km^2), and of those areas 53.5% and 16.29% respectively are influenced by dune stabilization. Continuing stabilization of the Athabasca sand dunes region may present conservation concerns for these narrowly distributed endemic taxa.

3.2 Introduction

The Athabasca sand dunes in northern Saskatchewan and northeast Alberta are the largest complex (~349 km^2) of active sand dunes in Canada and are one of the most northerly active sand dune formations on earth. This unique boreal landscape is characterized by large areas of active sand dunes and a unique cluster of nine narrowly distributed endemic plant species adapted to an environment of moving sand (Raup 1936, Argus and Steele 1979, Raup and Argus 1982, Macdonald et al. 1987, Cooper and Cass 2003, Lamb and Guedo 2012). These species are *Achillea millefolium* var. *megacephala* and *Tanacetum huronense* var. *floccosum* (both Asteraceae), *Armeria maritima* ssp. *interior* (Plumbaginaceae), *Deschampsia mackenzieana* (Poaceae), *Salix brachycarpa* var. *psammophila*, *Salix silicicola*, *Salix turnorii*, *Salix tyrrellii* (Salicaceae), and *Stellaria arenicola* (Caryophyllaceae). With the exceptions of *Salix tyrrellii* and *Stellaria arenicola* ("Not at Risk"), the Athabasca sand dunes endemic flora are currently listed as "Species of Special Concern" under the Canadian Species at Risk Act (SARA). These taxa were first described by Raup (1936), and while not all are currently recognized by Flora of North America Editorial Committee (1993+), these taxa are clearly morphologically and genetically distinct from locally co-occurring congeners (e.g. Purdy and Bayer (Purdy and Bayer 1996) and (Purdy and Bayer 1995)). Additionally, as the taxa are listed under SARA, updated status assessments are required regardless of taxonomic status. As sand dune systems are highly dynamic in nature, the extent, distribution pattern, and changes to the preferred habitat is a major knowledge gap for a comprehensive reassessment of the status of these species. We used remote sensing techniques to investigate morphologies and expansion process of sand dunes, long-term trends of vegetation establishment, and endemic plant habitat trends in the Athabasca sand dunes.

Species of conservation concern typically have relatively small populations. The sampling of such species results in small sample sizes influencing on the statistical power and model robustness (Stockwell and Peterson 2002, Thuiller et al. 2004, Guisan et al. 2006a, Pearson et al. 2007, Wisz et al. 2008). Relatively few studies have evaluated the effectiveness of different modelling algorithms with limited occurrence data for species specialized in dune environments (Wisz et al. 2008, Williams et al. 2009, Gogol-Prokurat 2011). We compared five different species distribution modelling algorithms generalized linear models (GLM), generalized additive models (GAM), multivariate adaptive regression splines (MARS), classification and regression trees

(CART), and Artificial neural networks (ANN)) with remotely sensed predictors to model the distribution of the Athabasca endemic plant species. Comprehensive field assessments of sand dune environments are limited by logistical considerations, particularly in remote sites such as the Athabasca sand dunes (Carson and MacLean 1986, Paisley et al. 1991, Wolfe et al. 2001, Okin and Painter 2004, Wood et al. 2012). The use of remotely sensed data and Global Information Systems (GIS) are an important part of sand dune environment analysis, as it enables studies across both greater spatial scales and more extended periods of time (Hugenholtz 2005b, a, Ewing and Kocurek 2010, Mohamed and Verstraeten 2012). The effectiveness of these remote sensing tools and approaches are well characterized, and use has been encouraged by low-cost availability of remotely sensed data (U.S. Geological Survey 2014).

Extensive research has been conducted on the Athabasca endemic flora exploring species morphological differences, taxonomic relationships, environmental/habitat affinities, ecological relationships, distribution patterns, and potential threats (Raup 1936, Argus and Steele 1979, Raup and Argus 1982, Macdonald et al. 1987, Cooper and Cass 2003, Lamb and Guedo 2012, Guy et al. 2013). The Athabasca sand dune endemic vascular plant species are highly adapted to an environment of moving sand and are rarely found in the forested landscapes around the dune fields (Raup and Argus 1982, Macdonald et al. 1987, Cooper and Cass 2003, Lamb and Guedo 2012). Evaluating regional-scale changes to the extent of the open sand environment and resulting changes in species distributions are critical to a long-term threats assessment of the Athabasca endemic flora. Our objectives are threefold: 1) to assess the population size and habitat extent for each endemic species, 2) better understand the dune environment by evaluating dune morphologies, long-term dune spatio-temporal variations, and 3) assess rates of woody vegetation encroachment and dune stabilization to evaluate an important potential threat to the Athabasca endemic flora.

3.3 Materials and methods

3.3.1 Study area

The Athabasca sand dunes in northern Saskatchewan and Alberta is the largest complex of active sand dunes in Canada (58°42" N ; 108°42" W) with a total extent of ~349 km^2 (Raup and Argus 1982, Carson and MacLean 1986). The largest dune field comprises two major areas on the

west (William River dune field) and east (Thompson Bay dune field) sides of the William River valley. A third main dune field (McFarlane River dune field) is on the west side of McFarlane River valley. A number of smaller dune fields are interspersed in between the major fields (Figure 3.1). Annual rainfall and the total precipitation including snow in the region are on average ~250 mm and ~370 mm, respectively (Government of Canada 2015). The climate is comparatively dry in the early part of the summer (May - June) and the beginning of fall (August - October). Total monthly rainfall ranges from ~20-70 mm during summer months (Government of Canada 2015). The highest average temperature of the region reaches ~18-20 ^{o}C in mid-summer (Government of Canada 2015).

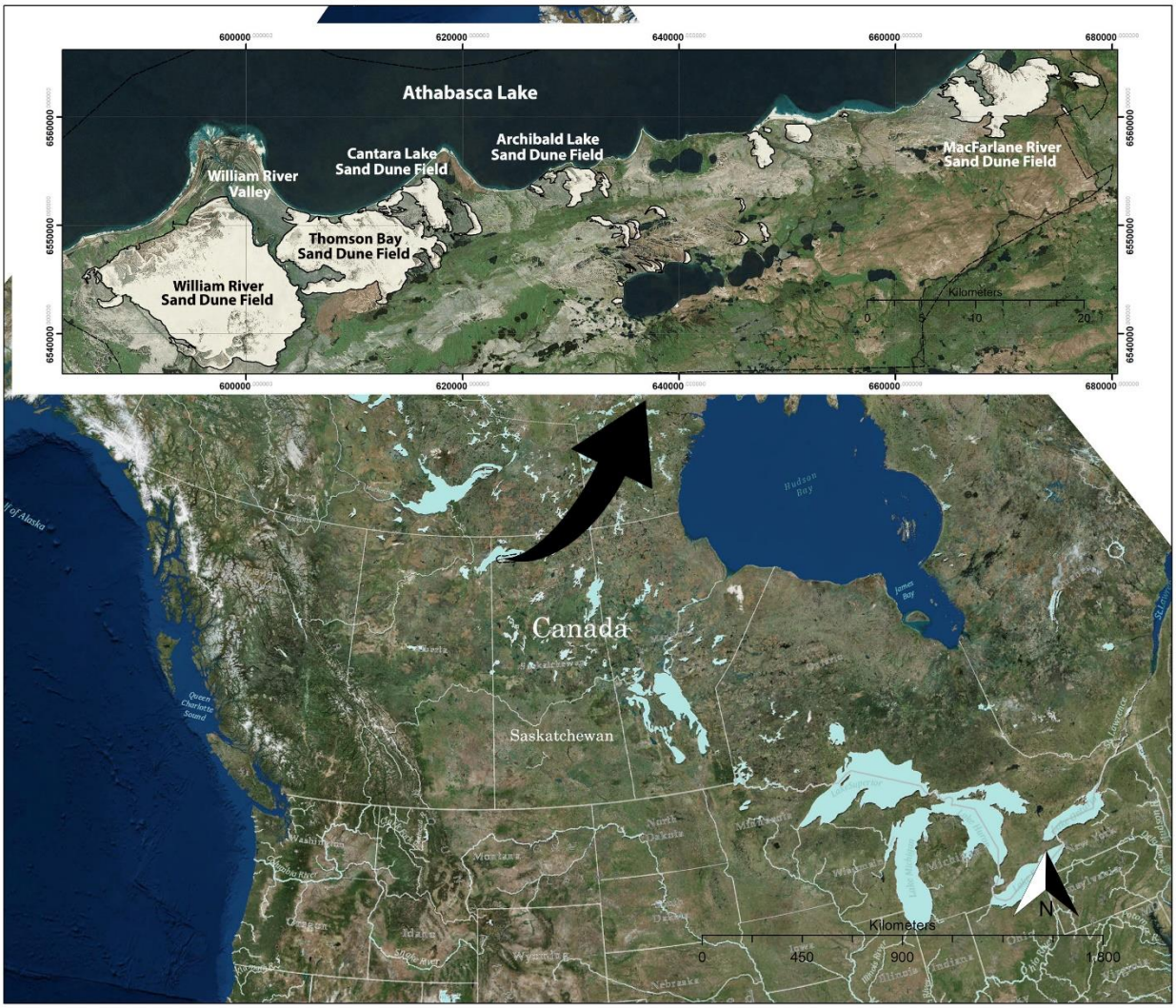


Figure 3.1: Study site. Landsat 5 TM true color composite image of Athabasca sand dunes in northern Saskatchewan acquired on June 14th, 2007.

3.3.2 Field data collection

Ground-truthed data for species and habitat used for the present study come from an extensive field survey conducted in 2009 and 2010 to assess populations distribution patterns, and the ecological relationships of the Athabasca endemics (Lamb et al. 2011, Lamb and Guedo 2012). The survey covered 224 randomly pre-located 250 m transects running east to west on constant northing (Lamb et al. 2011, Lamb and Guedo 2012). A detailed map of the transects and description of the allocation process available in Figure S3.1. All endemic species individual presence/absences of the target taxa were surveyed in a 10-m-wide transect for willow species and a 4-m-wide transect for grasses and forbs.

3.3.3 Satellite imagery and pre-processing

Two Landsat 5 TM images acquired on September 23rd, 2009 and July 8th, 2010 the closest possible date matches to field survey were used to develop species distribution maps (Referred hereafter as the 2009 and 2010 images). The 30 m spatial resolution of the Landsat images was not adequate to analyze sand dune dynamics on a seasonal or annual basis. Therefore, three image pairs covering the longest possible timespan were selected to assess long-term trends in the dune field extent. Images for multi-temporal environment change analysis were selected to have close Julian dates to minimize seasonal variation in vegetation phenology and reflectance characteristics (Jensen 2005). The selected images were acquired on July 3rd, 1985, July 10th, 2002, June 14th, 2007 and, 1st June, 2014 (Referred hereafter as 1985, 2002, 2007, and 2014 images). Sand dunes features (boundaries and crests) are spectrally distinct and less influenced by scattering in the near-infrared band (NIR) width (Paisley et al. 1991, Levin et al. 2004, Mohamed and Verstraeten 2012). The NIR band from all four images was used for sand dune migration and morphology analysis. Figure S3.2: Workflow overview describes all image analysis steps followed in order, images used and the intended uses of each step in the study. Conversion of Digital Number (DN) to radiance-at-sensor, then to ground reflectance was implemented at the beginning of data analysis following Chander et al. (2009). PCI Geomatica 2015 – Focus was used for atmospheric correction.

3.3.4 Imagery analysis methods

3.3.4.1 Overview of methods and objectives

Habitat occupancy modelling for endemic species was performed using five different modelling algorithms categorized under modern regression and machine learning techniques. Only the 2009 and 2010 images were used for this purpose and the likelihood of habitat occupancy of all nine endemic species were separately modeled using the most precise modelling technique selected from the training process. The Bi-Temporal Layer Stack (BTLS) technique was used to illustrate sand dune morphological features and spatio-temporal changes using the 1985 and 2014 image pair. The Post-classification Comparison Change Detection (PCCD) procedure was used to understand how sand dune encroachment occurs into surrounding vegetation and how vegetation encroachment occurs into sand dune fields. Rates of land cover change between sand and vegetation were estimated using 1985 as a base year for 2002, 2007, and 2014 images separately. The direction and the movement distance of sand dunes and surrounding vegetation at dune field boundaries were estimated using Generalized Additive Models (GAM). Finally, long-term temperature, wind and precipitation data from regional weather stations were examined to identify the climate influences on large-scale sand dune field changes. A graphical summary of the overall process is available in Figure S3.2, and details of all steps followed for each method are available in File S2 – Detailed Methods.

3.3.4.2 Athabasca endemic plant habitat modelling

The modelling began with observations of species presence/absence and identification of associated predictor variables likely to influence or describe the habitat and/or species occupancy. The training process links predictors with species presence/absence in geographical space to estimate the likelihood that the species is present or absent in un-sampled locations. Modelling techniques used in the study were from two groups: modern regression and machine learning algorithms as no studies have focused on modelling sparsely distributed species in dune habitats (Elith and Graham 2009, Elith and Leathwick 2009, Franklin and Miller 2009). Modern regression techniques tested include generalized linear models (GLM), generalized additive models (GAM), and multivariate adaptive regression splines (MARS). Machine learning algorithms tested include classification and regression trees (CART), and artificial neural networks (ANN).

We used measures that are independent of the event in the sample, commonly called as threshold-independent measures of accuracy to select the most suitable method for modelling target plant species occurrences (Franklin and Miller 2009). These include the "area under the curve" (AUC) of the receiver-operating characteristic (ROC) plot. ROC is a graph of the false positive error rate on the x-axis versus the true positive rate on the y-axis corresponding to each possible value of threshold probability; AUC is calculated by summing the area under the ROC curve where the value is >0.5 (performance better than random). AUC is a reliable measure for model comparisons as AUC is not affected by changes in species prevalence (Manel et al. 2001, Franklin and Miller 2009). The most precise modelling algorithm for each target species was selected from an iterative process with 1000 iterations for each species and algorithm combination. Average AUC of the 1000 iterations was calculated; the method with the highest average AUC was selected for final species occupancy modelling. Welch's One-way Analysis of Variances procedure and Games-Howell pair-wise mean comparison was used to illustrate differences and/or similarities of mean AUC among modelling algorithms for a given species (Sheskin 2003). Both Welch's ANOVA and the Games-Howell method are known to be robust where variances among factors are unequal (Sheskin 2003).

Predictive maps of each species were developed in ArcGIS Model Builder using the best model for the 2009 and 2010 analysis, and the location predictions were averaged for each species final prediction probability maps. Prediction accuracies were evaluated using 30% of ground truth data held back from the model training process. A series of threshold probabilities (0.5, 0.6, 0.7, 0.8, 0.9, 0.95 and 0.99) were used to illustrate changes in the error of commission, error of omission, overall accuracy and kappa statistic. Sand dune stabilization influences on available habitat area estimates were also assessed based on the above thresholds to identify the most appropriate thresholds in this context.

3.3.4.3 Analysis of sand dune morpho-dynamic characteristics

Analysis of sand dune morphological features and migration patterns were used to examine dune activity between 1985 and 2014. When the same spectral band from either three or two different dates is mounted to red, green, and blue colour guns (Write Function Memory banks), the change in reflectance is displayed in a different colour (Bi-Temporal Layer Stack; BTLS)(Mohamed and Verstraeten 2012). In instances where dune mineralogy is homogeneous,

<35% vegetation cover, and no biogenic-soil crust, changes in sand dune reflectance profiles are controlled mainly by illumination and shading effects from the orientation of the dune crest towards sun azimuth and elevation angles (Paisley et al. 1991, Levin et al. 2004, Okin and Painter 2004). Single dune and dune field scale analyses have shown that the most important feature for the study of sand dune morphological variation is the dune crest, as the crest always has high reflectance relative to other dune features in Landsat – NIR bandwidth (0.76-0.90 µm) (Paisley et al. 1991, Levin et al. 2004, Tsoar 2004, Ewing et al. 2006, Livingstone et al. 2007, Mohamed and Verstraeten 2012). Therefore, this study used dune crest and/or slip-face migration as a spectrally stable feature to identify sand dune morphological features and spatio-temporal variation.

3.3.4.4 Analysis of sand dune creation and vegetation encroachment

Post-classification Comparison Change Detection (PCCD) was used to measure both sand dune encroachment into surrounding forest vegetation and vegetation encroachment into stabilized sand dune fields. Comparisons were made using the 1985 image as the base year to 2002, 2007 and, 2014 images. An unsupervised classification approach was selected as *a-priori* detailed regional land-cover classes were not available. Twenty spectral classes were initially requested using the Iterative Self-Organizing Data (ISODATA) algorithm. Similar classes were merged to obtain three distinct land cover classes: water, vegetation and open sand. Next, the classified 1985 base year image was compared with a classified 2014 image on a pixel-by-pixel basis using a change detection matrix (Jensen 2005, Neigh et al. 2008). A from-to analysis was carried out to identify land cover changes from vegetation to sand and from sand to vegetation using Engineering Analysis and Scientific Interface (EASI) modelling in PCI Geomatica 2016 – Focus (Jensen 2005, Neigh et al. 2008).

The two techniques; BTLS and PCCD, however, were not able to clearly prove the sectors of the dune field boundaries where dune creation and vegetation encroachment had occurred. We, therefore, evaluated changes in reflectance values along the dune edges using 1985 as the base year to the years 2002, 2007 and, 2014. A series of 500 m long sampling transects crossing the dune edge were created every 1 km along the boundaries (Figure S3.3 and S3.4) and the reflectance values of each pixel underneath each transect were extracted to evaluate reflectance variations over time. Since sand has higher reflectance than vegetation, a positive reflectance difference indicates

dune migration into surrounding vegetation and negative differences indicates vegetation encroachment into sand dune fields.

Reflectance differences for each image pair (2002, 2007 and, 2014) to 1985 were modeled as a function of distance along the transects using Generalized Additive Models (GAM). GAM is a nonparametric extension of the Generalized Liner Model (GLM) and a convenient method to model nonlinear relationships when there is no prior expected shape to the curve (Hastie and Tibshirani 1990, Wood 2006). All statistical analyses were implemented in the R 3.1.2 software environment and the GAM was done using the "gam" function in the mgcv library (Wood 2006, R Core Team 2014). Directional categories were separately analyzed to identify predominant directions of sand dune expansion into vegetation and vegetation encroachment into sand dunes.

3.3.5 Analysis of climatic factors

In general, sand dune environments are influenced by long term wind and rainfall patterns (Carson and MacLean 1986, Wolfe et al. 2001, Tsoar 2004, Hugenholtz 2005b, a, Tsoar 2005, Hugenholtz et al. 2012, Wood et al. 2012, Al-Masrahy and Mountney 2013). Long term climatic data for the study were obtained from the Environment Canada - Fort Chipewyan (58°46" N ; 111°07"' W) weather station located approximately 115 km southeast of Athabasca dune fields, as this was the closest station with long term weather data available (Carson and MacLean 1986). Monthly directional wind patterns were examined by summarizing hourly wind direction readings separately by month from 1971 to 2015. Rainfall, snowfall, total precipitation and temperature data were examined from 1967 to 2006 (changed timespan based on data availability). Mean monthly total rainfall, mean monthly total snowfall, mean monthly total precipitation, mean monthly temperature, total annual rainfall and, total annual precipitation were analyzed to understand how seasonal variation may influence wind action on exposed sand.

3.4 Results

3.4.1 Athabasca endemic plant habitat modelling

The mean AUC for all species, analytical algorithm, and year combinations were above 0.5 (Figure 3.2), indicating that all model algorithm performances were better than random and capable of successfully modelling each species. Welch's One-way Analysis of Variance procedure

confirmed ($p<0.05$) significant differences in performance among algorithms (Table 3.1). The 2009 Games-Howell pair-wise mean comparison showed that GLM mean AUC was significantly higher for all species with the exception of *Deschampsia mackenzieana* (GAM highest AUC but GLM and GAM mean AUC's not significantly different). The 2010 data had GLM performance as the highest except for *Deschampsia mackenzieana*, *Stellaria arenicola*, and *Salix turnorii* (RF classification algorithm). The best algorithm for each species and year combination was used to develop a prediction map (Figure 3.3).

Table 3.1: Mean area under the curve (AUC) comparison among different modelling algorithms for each species.

The Analysis of Variance (ANOVA) procedure for mean comparison purpose tested the null hypothesis of all means are equal, against the alternative hypothesis of at least one mean is different at significance level 0.05. Welch's test was used in this analysis as equal variances were not assumed among treatment levels. The mean comparison was assessed using Games-Howell Pairwise Mean Comparisons. Means that do not share a letter are significantly different among modelling techniques along with each species.

Plant Species and Year of Field Survey		**Welch's Test – P value**	**Games-Howell Pairwise Mean AUC Comparisons and Grouping**				
			GLM	**GAM**	**MARS**	**RF**	**ANN**
2009	TANHUR	<0.00001	**0.70440** ^{A}	0.67428 ^{C}	0.68464 ^{B}	0.70071 ^{A}	0.68004 ^{B}
	STEARE	<0.00001	**0.76347** ^{A}	0.78582 ^{D}	0.75228 ^{B}	0.72050 ^{D}	0.74095 ^{B}
	SALTYR	<0.00001	**0.88121** ^{A}	0.76492 ^{D}	0.84582 ^{B}	0.83111 ^{C}	0.84135 ^{B}
	SALTUR	<0.00001	**0.83044** ^{A}	0.76468 ^{D}	0.81707 ^{B}	0.76509 ^{D}	0.78389 ^{C}
	SALSIL	<0.00001	**0.80436** ^{A}	0.78830 ^{B}	0.78859 ^{B}	0.74952 ^{C}	0.79146 ^{B}
	SALBRA	<0.00001	**0.84731** ^{A}	0.73274 ^{D}	0.83385 ^{B}	0.80933 ^{C}	0.81177 ^{C}
	DESMAC	<0.00001	**0.71887** ^{A}	**0.72370** ^{A}	0.71331 ^{B}	0.69717 ^{D}	069501 ^{D}
	ARMMAR	<0.00001	**0.74262** ^{A}	0.69189 ^{B}	0.69055 ^{B}	0.67615 ^{C}	0.67615 ^{A}
	ACHMIL	<0.00001	**0.87346** ^{A}	0.66640 ^{E}	0.82711 ^{D}	0.85904 ^{B}	0.84459 ^{C}
2010	TANHUR	<0.00001	0.65215 ^{B}	0.63649 ^{C}	0.64785 ^{B}	**0.70591** ^{A}	0.64992 ^{B}
	STEARE	<0.00001	0.59273 ^{E}	0.66997 ^{B}	0.64304 ^{C}	**0.69059** ^{A}	0.60836 ^{D}
	SALTYR	<0.00001	**0.72060** ^{A}	0.69598 ^{C}	0.71238 ^{B}	0.71069 ^{B}	0.71248 ^{B}
	SALTUR	<0.00001	0.69428 ^{B}	0.69289 ^{B}	0.68453 ^{C}	**0.71500** ^{A}	0.65775 ^{D}
	SALSIL	<0.00001	**0.71709** ^{A}	0.68501 ^{C}	0.69441 ^{B}	0.67657 ^{D}	0.69348 ^{B}
	SALBRA	<0.00001	**0.73145** ^{A}	0.71409 ^{C}	0.72411 ^{B}	0.70143 ^{D}	0.71250 ^{C}
	DESMAC	<0.00001	**0.752754** ^{A}	0.749688 ^{A}	0.739501 ^{B}	0.712835 ^{C}	0.742380 ^{B}
	ARMMAR	<0.00001	**0.68492** ^{A}	0.61021 ^{C}	0.58701 ^{C}	0.55524 ^{D}	0.62506 ^{B}
	ACHMIL	<0.00001	**0.76801** ^{A}	0.64313 ^{C}	0.74405 ^{B}	0.74057 ^{B}	0.64192 ^{C}

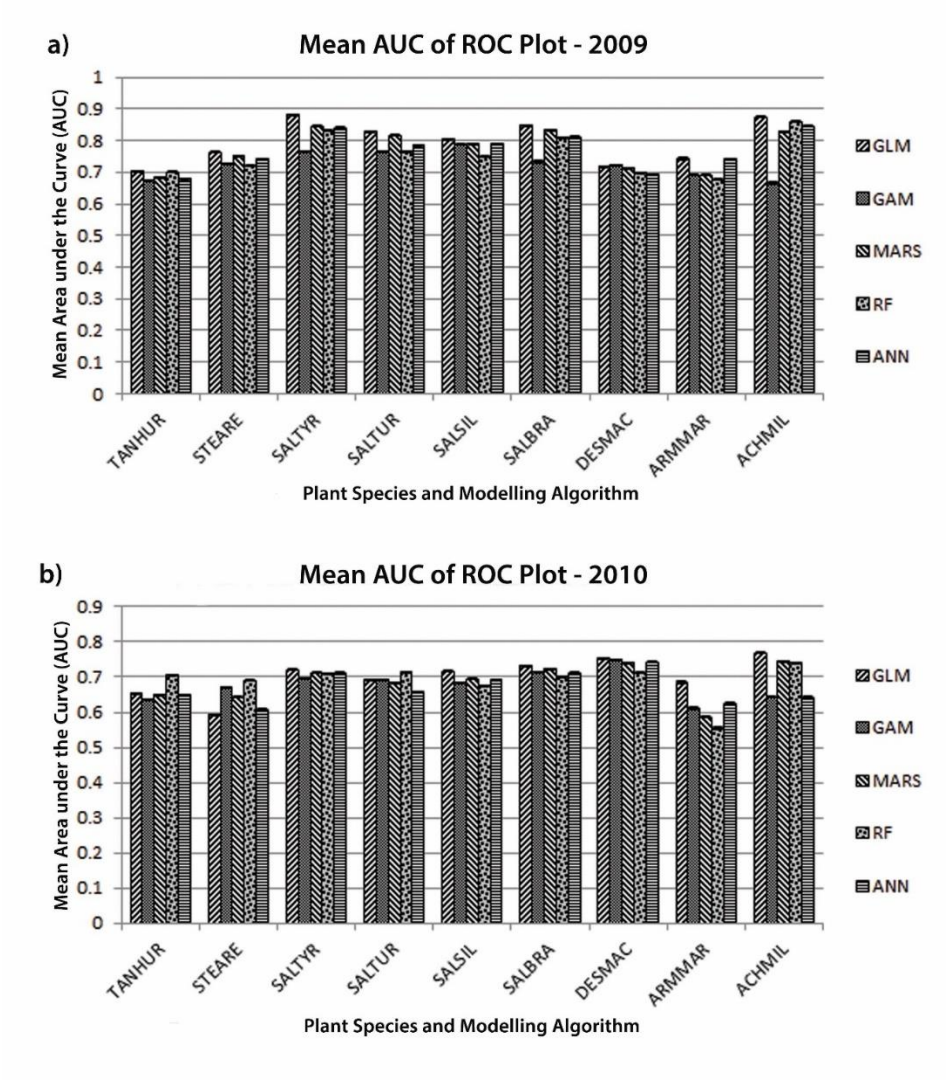


Figure 3.2: Modeling algorithm performance evaluation based on the Area Under the Curve (AUC) of Receiver Operating Characteristic Plot (ROC). The Y-axis is the mean AUC of the ROC value for each species and each modeling algorithm. Each bar for a species represents the calculated mean AUC of one thousand iterations of a particular algorithm. Species abbreviations refer to Tanacetum huronense var. floccosum (TANHUR), Stellaria arenicola (STEARE), Salix tyrrellii (SALTYR), Salix turnorii (SALTUR), Salix silicicola (SALSIL), Salix brachycarpa var. psammophila (SALBRA), Deschampsia mackenzieana (DESMAC), Armeria maritima ssp. Interior (ARMMAR), and Achillea millefolium var. megacephala (ACHMIL). Modeling techniques include Generalized Linear Models (GLM), Generalized Additive Models (GAM), Multivariate Adaptive Regression Splines (MARS), Random Forest (RF), and Artificial neural networks (ANN).

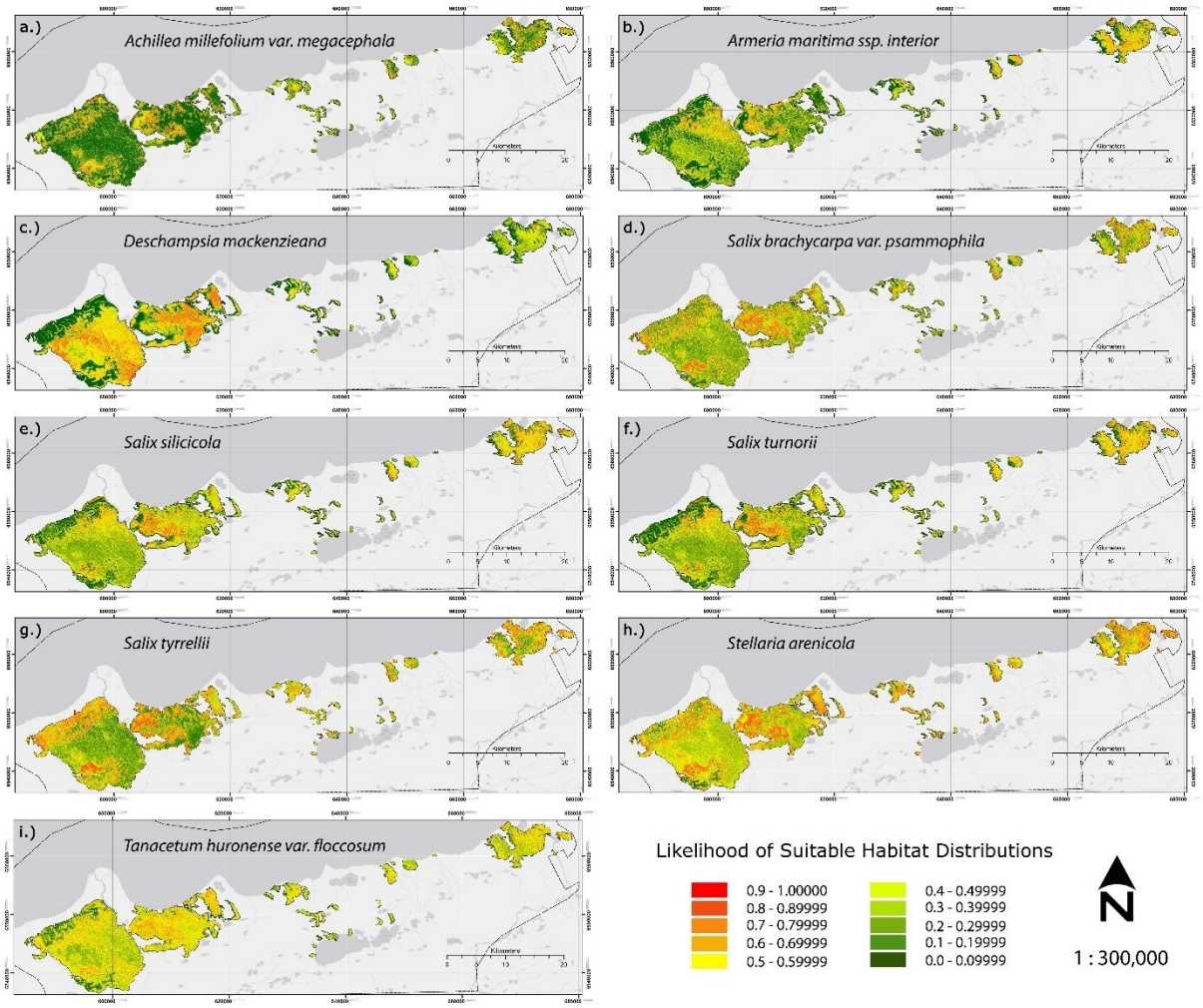


Figure 3.3: Predicted likelihood of suitable habitat distributions for Athabasca endemic plant species. Probability of suitable habitat within a pixel for each of the Athabasca endemic species. The color ramp uses ten categories with 0.1 increments. Warm colors indicate higher probability and the cooler colors lower probability of finding suitable habitat.

We estimated the influence of sand dune stabilization on the extent predictions for each species over a range of threshold probabilities to demonstrate the assessment uncertainty associated with predictions at each level of threshold. Errors of commission, omission, overall accuracy and kappa statistic were evaluated and the result shows optimum levels are between 0.5 to 0.7 threshold probability range for species in consideration (Figure S3.5). The estimated total of available/occupied habitat and the extent of the stabilized/affected proportion of the habitat

gradually decreased with increases in the threshold. Slope variation of estimated total occupied habitat extent in relation to increase in threshold were consistent (Lower extent estimates for least abundant species and vice versa) below 0.6 threshold and the prediction patterns are largely random above that point (Figure 3.4 and Table S3.2). For example, the estimated extent of the rare *Achillea millefolium* and widespread *Deschampsia mackenzieana* were 38.92 km^2 and 89.52 km^2 respectively at a 0.6 threshold, in comparison to 17.12 km^2 and 9.52 km^2 at the 0.8 threshold. More details of total occupied habitat extent estimate (km^2), stabilized habitat extent between 1985 to 2014 (km^2) and percent of most probable habitat influenced by sand dune stabilization based on varying thresholds is available in Table S3.2.

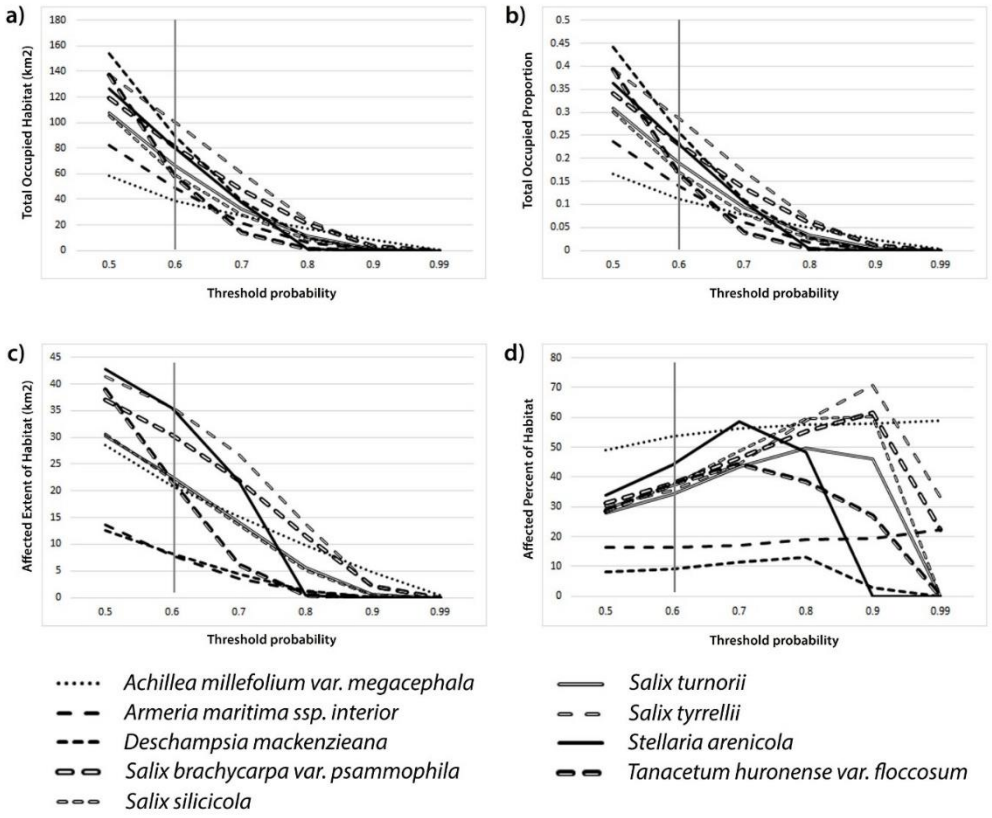


Figure 3.4: Analysis of estimate uncertainty based on varying threshold probabilities. a.) Total occupied habitat extent, b.) Occupied proportion in relation to total dune area, c.) Stabilized habitat extent between 1985 and 2014, and d.) Percent of most probable habitat influenced by sand dune stabilization. Each line is a representation of each species and estimate variations are based on various threshold probabilities. The line at the 0.6 threshold represents the most stable prediction in-comparison to observed prevalence patterns of species in consideration.

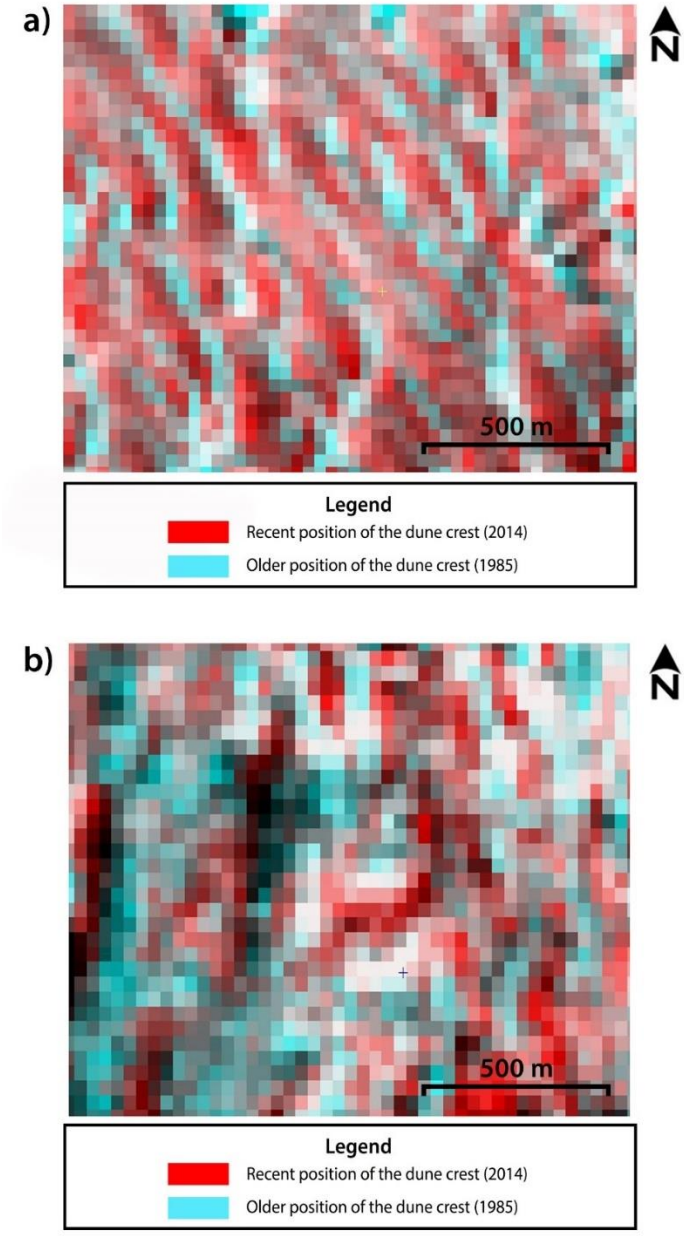


Figure 3.5: Subset of Bi-Temporal Layer Stack (BTLS) results. The red color areas indicate dune crest location of the recent image (increased reflectance over time) and the cyan color indicates dune crest location of the older image (decreased reflectance over time), a) The simple crescentic ridge-transverse dunes observed in McFarlane River dune field, b) The compound crescentic ridge-transverse dunes observed in the William River dune field.

3.4.2 Sand dune morpho-dynamic characteristics

The BTLS analysis revealed that the crests of the sand dunes were generally aligned in a northwest-southeast direction with dune structures were approximately straight or gently curved along the main axis (Figure 3.5a). The Athabasca sand dune region is comparatively dry (Figure 3.6) throughout the summer (May to July) and the beginning of fall (August - October) suggesting that winds during this period may dominate dune field migration patterns. The late-summer – early fall period is dominated by winds from the southeast and east (Figure 3.7). However, winds throughout the summer and early fall from the northwest to southwest range may influence longitudinal sand movement along the dune axis (Figure 3.7). The BTLS results and wind data demonstrates that the dune migration direction is almost parallel and the main axis of the dunes perpendicular to the prevailing wind pattern. The majority of dunes can thus be classified as "crescentic ridge" in external morphological terms and "transverse" in morpho-dynamic terms. Complex red-cyan patterns observed at the ends of individual sand dunes and several large areas in the William River and Thompson Bay dune fields indicate less distinct spectral profiles (Figure 3.5b) arising from gently undulating to flat sand surfaces.

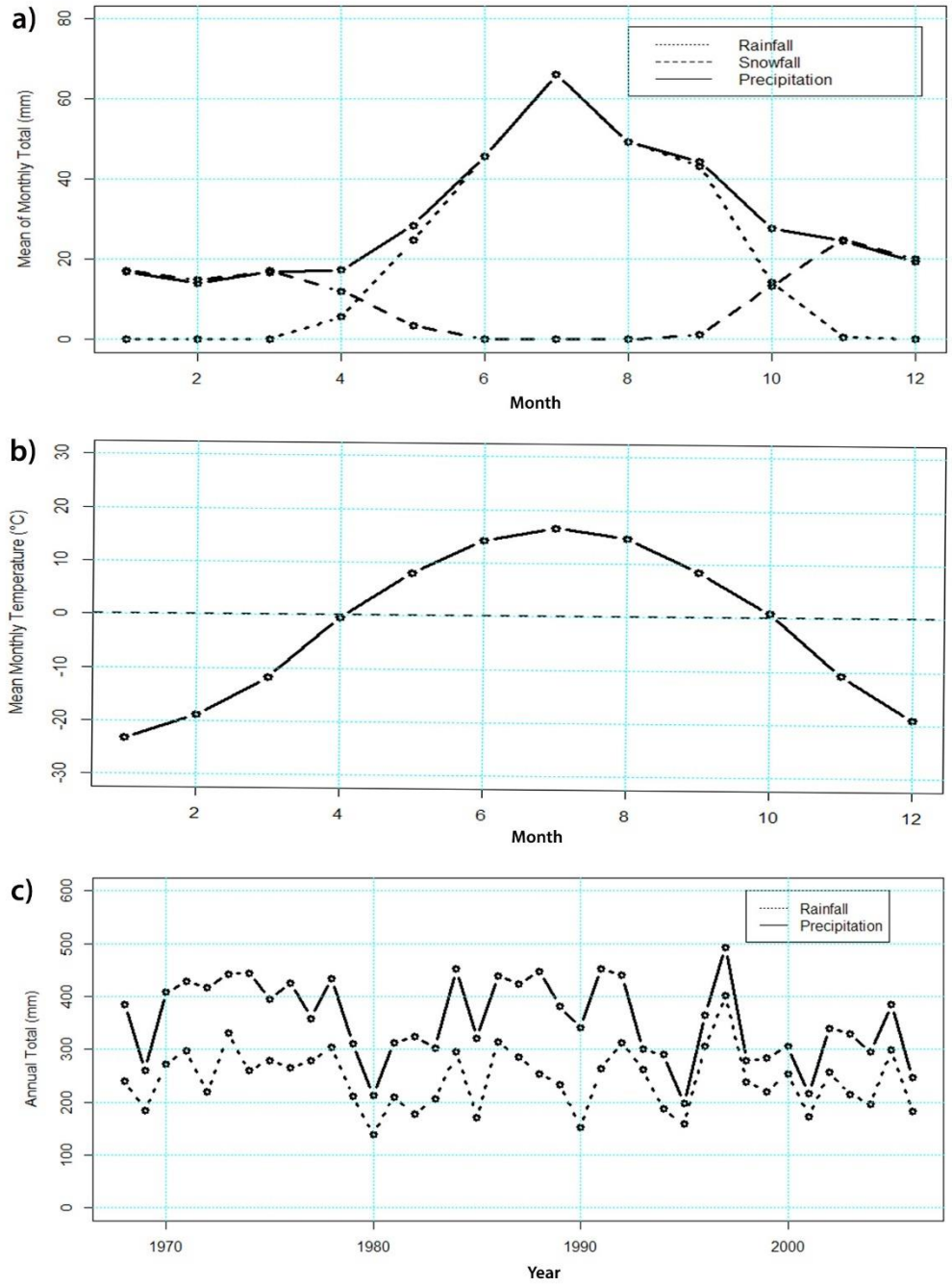


Figure 3.6: Rainfall, snowfall, total precipitation and, temperature. Daily readings of total rainfall and average temperature from the Environment Canada - Fort Chipewyan (58°46" N ; 111°07"' W) weather station located approximately 115 km southeast of Athabasca dune fields. The data were averaged across 1967 to 2006 to obtain monthly values. a.) Average monthly total rainfall, snowfall and, total precipitation. b.) Average monthly temperature. c.) Total annual rainfall and total precipitation.

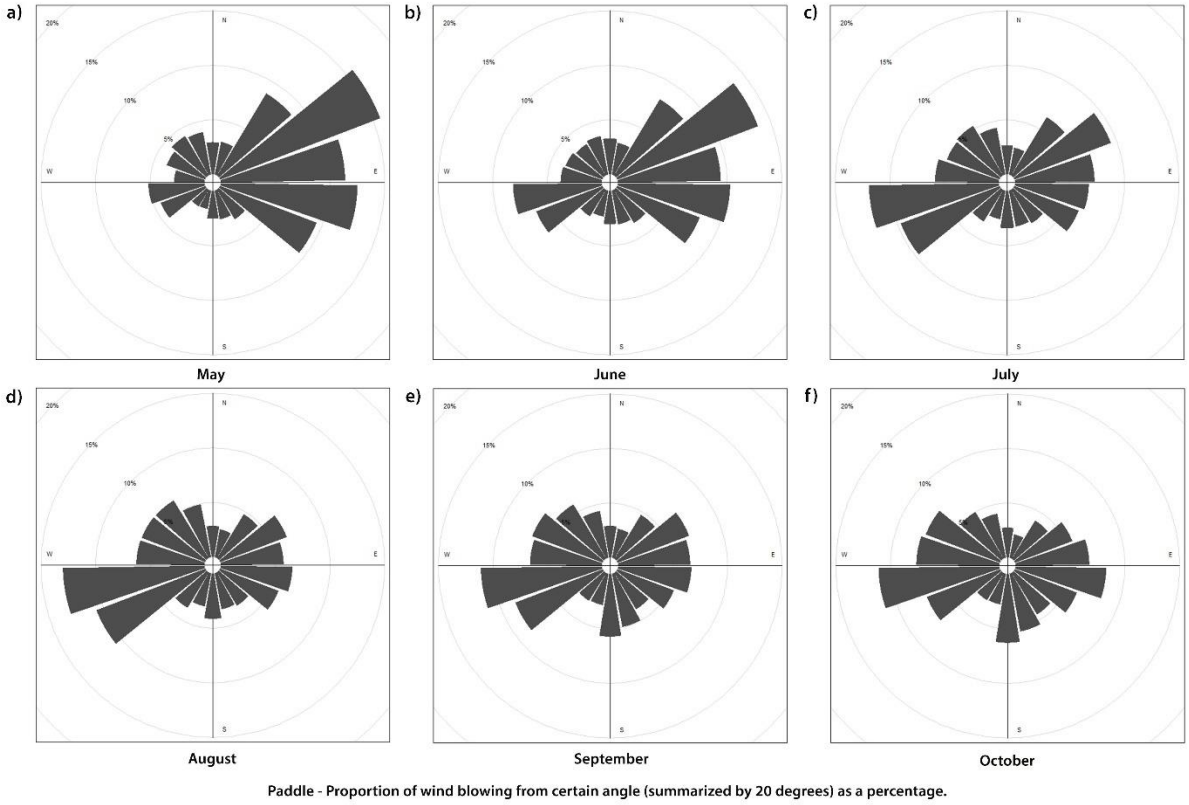


Figure 3.7: Monthly directional variations of wind pattern. Summary of hourly readings of wind direction obtained from Environment Canada - Fort Chipewyan (58°46" N ; 111°07'" W) weather station located approximately 115 km southeast of Athabasca dune fields. The true direction from which the wind is blowing measured in 10s of degrees. The wind rose plot summarizes the direction by 20-degree increments and each paddle represents proportion of wind observations from that angle. Measurements were recorded through 360 degrees and a calm wind is recorded as 0 degrees. The frequency variations of wind direction were analyzed on a monthly basis from 1971 to 2015. Only wind direction data (no speed data) were available for the time period reported.

3.4.3 Sand dune creation and vegetation encroachment

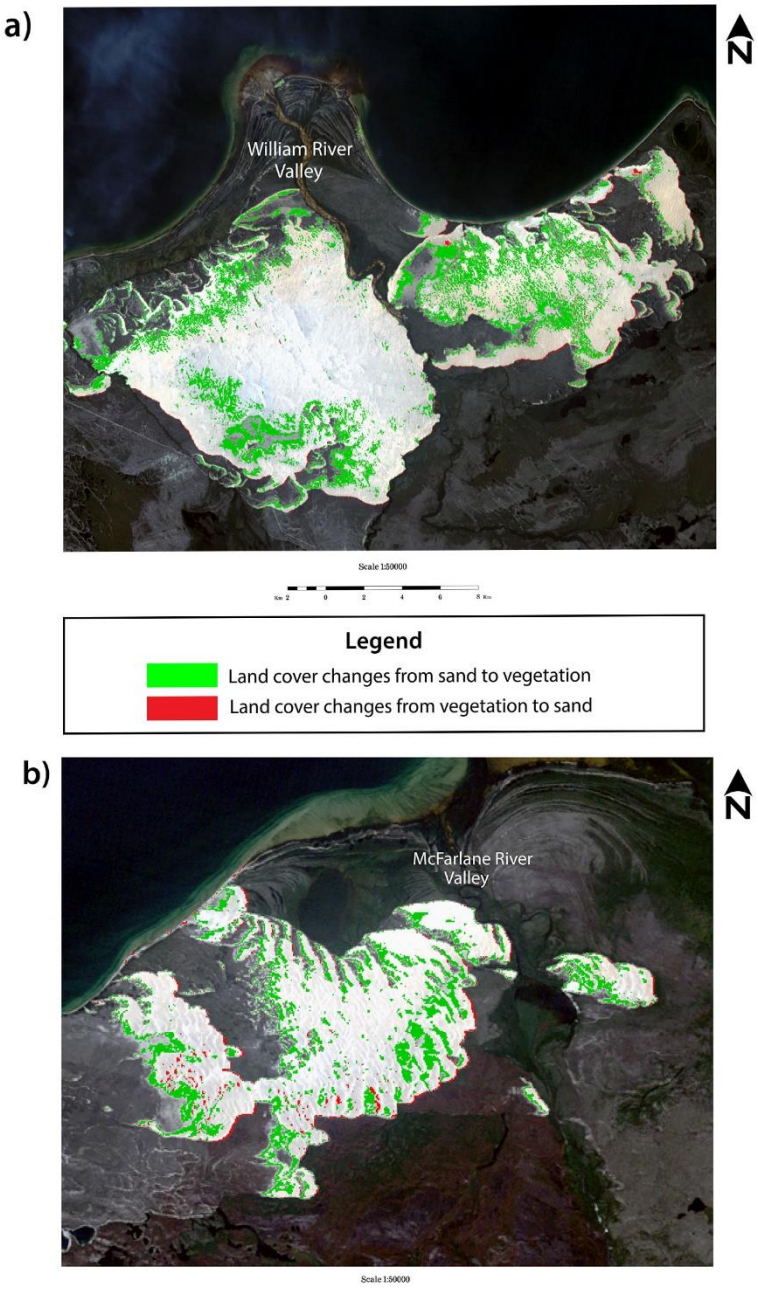


Figure 3.8: Post-classification Comparison Change Detection (PCCD) maps. The 1985 true color image was used as the base map to overlay PCCD results. The red color indicates land cover changes from vegetation to sand and the green color indicates land cover changes from vegetation to sand and the green color indicates land cover changes from sand to vegetation from 1985 to 2014. a) William River and Thompson Bay dune fields, and b) McFarlane River dune field.

The PCCD results indicate that changes from vegetation to sand (red) during the period of 1985 to 2014 were most prominent on the east and southeast margins of each sand dune field (Figure 3.8a and 3.8b), with the exception of the east edge of William River dune field where it borders the river. Overall, this indicates that sand dune creation occurs mainly at the east and southeast edges of each dune field. The estimated area of sandy surfaces created between 1985 and 2014 was 4.20 km^2 and the rate of change was 0.14 km^2 per year (Table 3.2). In contrast, green, indicating land cover changes from sand to vegetation during the period of 1985 to 2014, is more prominent on the west edge indicating a general pattern of forest encroachment on that edge (Figure 3.8a and 3.8b). The estimated area of sandy surface occupied by woody vegetation was 57.55 km^2 and the rate of change was 1.98 km^2 per year (Table 3.2). The rate of vegetation encroachment into the sand dunes is thus approximately 9 times higher than new sand dune creation resulting in a net loss of dune area from 1985 to 2014 of 53.76 km^2. Annual rates are variable as the estimated average annual rate of dune area lost from 1985 to 2002 was 1.15 km^2 per year, but 1985 to 2014 estimates were an average of 1.85 km^2 lost per year (Table 3.3).

Table 3.2: Post-Classification Comparison Change Detection (PCCD) estimate of total sand dune creation and dune stabilization from 1985 to 2002, 2007, and 2014.

The table illustrates the total sand dune creation and stabilization estimates of Post-Classification Change Detection process. Number of pixels was converted to square kilometers using 30 m spatial resolution of Landsat images used. The net creation or reduction of the dune area for each year was an estimate in comparison to 1985 base year. The rate of reduction was calculated considering the time gap between 1985 and the recent year in consideration.

Duration	**Number of Years**	**Number of Pixels**	**Extent of Vegetation Encroachment (km^2)**	**Rate of Vegetation Encroachment (km^2 $year^{-1}$)**	**Number of Pixels**	**Extent of Sand Dune Creation (km^2)**	**Rate of Sand Dune Creation (km^2 $year^{-1}$)**
1985-2002	17	26280	23.65	1.39	4145	3.73	0.22
1985-2007	22	31963	28.77	1.31	3681	3.31	0.15
1985-2014	29	63945	57.55	1.98	4670	4.20	0.14
Marginal Means			**36.66**	**1.56**		**3.75**	**0.17**

Table 3.3: Total sand dune net loss and the rate of change estimate from 1985 to 2002, 2007, and 2014.

The table illustrates the total open sand dune area of each year classified using unsupervised classification technique. Number of pixels was converted to square kilometers using 30 m spatial resolution of Landsat images used. The net reduction of the total dune area for each year was an estimate in comparison to 1985 base year. The rate of reduction was calculated considering the time gap between 1985 and the recent year in consideration.

Year	**Number of Pixels**	**km^2**	**Reduction of the Dune Area (km^2)**	**Number of Years**	**Rate of Reduction (km^2 $year^{-1}$)**
1985	293327	263.99	NA	NA	NA
2002	271667	244.50	19.49	17	1.15
2007	265027	238.52	25.47	22	1.16
2014	233595	210.24	53.76	29	1.85

GAM by directional category was used to estimate the distance and directional movement of sand dunes and vegetation at the dune boundaries. Overall, results were in line with the PCCD analysis with significant positive reflectance differences the south-east and east directions across all three comparisons from 1985 to 2002, 2007 and, 2014 (Table 3.4). This indicates that creation of sand dunes occurs most commonly on the east and southeast dune margins at a rate of 3.28 m $year^{-1}$ (Table 3.5, panel d and h of Figure S3.6, Figure S3.7 and, Figure S3.8). Negative reflectance differences indicating vegetation encroachment were significant on the western dune boundaries indicating encroachment at a rate of 5.85 m $year^{-1}$ across all three comparisons (Table 3.4, Table 3.5, panel a of Figure S3.6, Figure S3.7 and, Figure S3.8).

Table 3.4: Generalized Additive Modelling (GAM) results of directional movement and distance analysis of sand dune creation and vegetation encroachment in the study area.

Generalized Additive Modelling (GAM) by directional category was assessed to identify substantial changes in reflectance at boundaries of sand dune fields. Comparisons were made from 1985 to 2002, 2007, and 2014 and all models were assessed at 0.05 significance level. The table contains A) effective degrees of freedom, B) F-value, and C) p-value of the GAM process for each directional category. The significant positive reflectance differences (dune creation) towards the south-east and east directions were observed across all three comparisons from 1985 to 2002, 2007 and, 2014. The negative reflectance differences indicating vegetation encroachment were only significant on the western dune boundaries across all three comparisons.

		E-W	**N-S**	**NE-SW**	**NW-SE**	**S-N**	**SE-NW**	**SW-NE**	**W-E**
1985-2002	A	4.564	1.001	1.001	3.903	1.001	2.162	1.000	5.148
	B	16.534	8.559	8.113	9.144	9.332	8.451	10.420	18.040
	C	**<0.001**	**0.003**	**0.004**	**<0.001**	**0.002**	**<0.001**	**0.001**	**<0.001**
1985-2007	A	6.120	2.257	1.002	7.483	1.002	1.524	1.001	7.326
	B	15.210	2.778	2.407	9.202	0.782	1.417	1.265	13.597
	C	**<0.001**	0.043	0.120	**<0.001**	0.377	0.238	0.260	**<0.001**
1985-2014	A	6.443	5.604	1.003	7.492	1.001	2.068	1.000	7.596
	B	19.031	4.670	0.000	10.367	1.542	2.130	31.246	21.030
	C	**<0.001**	**<0.001**	0.997	**<0.001**	0.214	0.103	**<0.001**	**<0.001**

Table 3.5: The distance of sand dune creation (east and southeast) and woody vegetation encroachment (west) calculation from 1985 to 2002, 2007 and, 2014.

The table illustrates calculated distance of positive or negative reflectance difference was observed using Generalized Additive Modelling (GAM) technique for each pairwise comparison. The positive reflectance difference relationship indicates sand dune creation and the negative reflectance difference relationship indicates dense vegetation encroachment. The positive distance was calculated from zero (edge of the dune field) to a point where lower confidence limit crosses the main X-axis. The negative distance was calculated from zero (edge of the dune field) to a point where upper confidence limit crosses the main X-axis. More detailed illustration of distance calculation available in Figure S3.9. The calculation process was similar in each pairwise comparison and the marginal mean was calculated to obtain an average of all three estimates.

Duration	**Number of Years**	**Sand Dune Creation (m year^{-1})**	**Vegetation Encroachment (m year^{-1})**
1985-2002	17	3.49	7.11
1985-2007	22	3.79	5.66
1985-2014	29	2.57	4.77
Marginal Means		**3.28**	**5.85**

3.5 Discussion

3.5.1 Analysis of species occurrence likelihood of Athabasca endemics

We modelled endemic species habitat distributions to identify the likelihood of habitat occupancy of nine endemic vascular plant species observed in Athabasca sand dunes. The effectiveness of five algorithms was evaluated in light of the sparse distribution of the species and low spatial resolution of predictors. All methods showed robust performance (>0.5 AUC), however, strong GLM performance likely arises from strong linear relationships between species habitat occupancy and Landsat 7 reflectance bands. The literature highlights that habitat specialists are often more accurately modelled than generalists because of their reliance on spatially restricted and spectrally distinct environmental conditions(Guisan et al. 2006a, Franklin and Miller 2009). However, we were constrained in this analysis by the 30m spatial resolution and strongly believe that the observed uncertainties may have been avoided with high-resolution imagery. This applies most strongly for *Armeria maritima* and *Achillea millefolium* as they are the least abundant of the species and are restricted respectively to limited habitats in gravel pavements and wet inter-dune slacks scattered among the active dunes (Lamb and Guedo 2012).

Analysis of available/occupied habitat extent in relation to varying probability thresholds is a good illustration of how habitat predictions can be influenced by threshold probabilities. All species had a decreasing trend of estimated occupied habitat with increasing threshold probability. The rare habitat specialist species *Achillea millefolium* and *Armeria maritima* had very low estimated occupied proportion relative to common species such as *Deschampsia mackenzieana* that had a large observed habitat extent. A similar trend of estimated relative abundances was observed between the 0.5 to 0.6 threshold levels, however the pattern of relative abundances drastically deviates from observed pattern beyond the threshold 0.6. Based on this we used a 0.6 threshold as our cut-off probability to illustrate sand dune stabilization influences on the habit for each species.

Achillea millefolium habitat was the most influenced by sand dune stabilization with an estimated total available/occupied habitat of 38.92 km^2 (11.14% of total dune area) and 53.5% of that area potentially influenced by stabilization. *Armeria maritima* had 48.82 km^2 (13.97% of dune area) of available/occupied habitat and but only 16.29% potentially influenced by dune

stabilization. The habitat generalist *Deschampsia mackenzieana* in contrast occupies 89.52 km^2 (25.63% of the dune area) with only 9.18 % of that area influenced by stabilization. *Salix brachycarpa, Salix silicicola, Salix turnorii, Salix tyrrellii, Stellaria arenicola* and *Tanacetum huronense* respectively had 37.78, 37.22, 34.21, 35.58, 44.20 and 37.55 percent of occupied habitat potentially affected by stabilization.

The risk of stabilization may be of importance to the long-term viability of the populations of the less abundant endemic species, particularly *Achillea millefolium* (Lamb et al. 2011, Lamb and Guedo 2012). *Armeria maritima* is found in very low numbers only on gravel pavements and occasionally in wet inter-dune slacks; the species is absent from active dunes as it cannot tolerate burial (Lamb and Guedo 2012). Gravel pavements are generally found near the center of the dune fields, however, and are thus at low risk of stabilization. *Achillea millefolium* is potentially more vulnerable to stabilization. It was the second least abundant species observed during the field survey and is frequently found in wet inter-dune slacks and secondarily on lichen crowberry heaths and woodlands near dune edges (Lamb and Guedo 2012). The presence of *Achilliea* habitat near the dune margins likely drives the very high percentage of habitat potentially affected by stabilization, and suggests that conservation concern for this species may be warranted.

Wide habitat preferences ensure the availability of abundant suitable habitat for the more common endemic species (*Deschampsia mackenzieana, Salix brachycarpa, Salix silicicola, Salix turnorii, Salix tyrrellii, Stellaria arenicola*, and *Tanacetum huronense*). That, combined with the relatively large populations (Lamb and Guedo 2012), suggests low conservation concern for those species. The extent of the wet inter-dune slacks remains an important question, however, as that habitat is a site of high recruitment for *Salix brachycarpa, Salix silicicola, Salix turnorii, Salix tyrrellii, Achillea millefolium, and Tanacetum huronense* (Lamb and Guedo 2012). A long-term understanding of the distribution of this habitat is important as it supports higher endemic species richness and total abundance than anywhere else on the landscape.

3.5.2 Sand dune morpho-dynamic characteristics

We observed many transverse-crescentic ridges in the McFarlane river area (Figure 3.5a) migrating in a northeasterly direction. The crescentic ridge classification is based on the external morphology of depositional forms (shape) of the dunes and is typically straight or very gently

curved at both ends. The formation of crescentic dunes starts from individual crescents that coalesce laterally when deposition of sand increases. Transverse dunes imply a morpho-dynamic process driven by the wind with sand movement and dune migration towards the prominent wind direction and the main dune axes perpendicular to the wind direction (McKee 1979, Lancaster 1995, Mohamed and Verstraeten 2012). Carson and MacLean (1986), made similar findings in the Athabasca dunes system, noting the abundance of transverse sand dunes migrating towards the northeast. Furthermore, they identified longitudinal processes that were likely driving sand dune lateral coalition and elongation. Although crescentic ridge features are very clearly visible at the center of single dunes in the BTLS maps, the features are compound at both ends reflecting lateral coalition processes were active in the past. Frequent winds from the west-northwest and east-southeast in July – October (Figure 3.7) may contribute significantly to the lateral coalition and elongation processes.

Compound dune morphologies are common in the William River and Thompson Bay dune fields (Figure 3.5b). The very complex red and cyan pixel patterns make the distinction of simple morphological variation using the BTLS procedure difficult. However, positional changes in high reflectance areas are an indication of dune crest changes over time. The fieldwork of Carson and MacLean (1986), confirms that the area was composed of transverse crescentic ridges and gently undulating to flat sand surfaces. This can result in similar reflectance profiles throughout the area which are difficult to distinguish at the 30 m spatial resolution of Landsat images. According to Lancaster (1995), compound crescentic ridges are characterized by superimposed multiple crescentic ridges on the upper stoss and crestal areas of the major crescentic ridge. This is most likely the reason for the complex red-cyan pixel patterns as crescentic ridges are formed by individual crescents coalescing laterally when deposition of sand increases. Furthermore, analysis of compound crescentic ridge dune morphology is challenged by the presence of similar reflectance profiles of flat sand sheets in inter-dune areas. The compound morphology tends to create less distinct spectral patterns in comparison with simple dune morphologies. Overall, the Athabasca dune system was likely dominated by the movement of sand in the 45 to 135 degree (northeast to southeast) directional range.

3.5.3 Sand dune creation and vegetation encroachment

The main zones of sand movement into existing vegetation identified in the PCCD and GAM approaches were along the east and southeast edges of the sand dune fields. The rate of creation of new sand surfaces was 0.14 km^2 $year^{-1}$, a very slow process in comparison to total dune area (~349 km^2) and open sand surface (~225 km^2). Similarly, the GAM shows that positive reflectance difference moved at approximately 3.28 m $year^{-1}$ only on the southeast and east edges of the dune fields. The analysis of climate data supports both the PCCD and GAM results regarding the influences of the wind regime on dune creation and vegetation encroachment. Winds towards the southeast to northeast directional range (winds from 220° to 310° directional range) were most common in the area at the end of the summer and the beginning of the fall (Figure 3.7) in combination with low precipitation (Figure 3.6) reflects the most potential drive to move sand. Carson and MacLean (1986), stated that the Athabasca sand dunes were migrating to the northeast at a much lower rate of about 0.5 m $year^{-1}$. This difference likely reflects the spatial and temporal limitations of short-term field-based work to measure a slow process dependent on variable weather conditions.

Vegetation growth into the dunes between 1985 and 2014 mainly occurred at the west sides of the sand dune fields. The total area occupied by woody vegetation from 1985 to 2014 was 57.55 km^2 and the rate of change was 1.98 km^2 $year^{-1}$. This likely reflects reduced sand dune activity on the west side of dunes as the predominant sand dune movement was to the east. The GAM shows that the negative reflectance difference moved approximately 5.85 m $year^{-1}$. The exact mechanisms driving reductions in active sand movement on the west side of dune fields are unclear. However, higher frequencies of wind towards the east and southeast directions in the summer may contribute to active sand movement opposite to west boundaries of the dune system. Coniferous tree spread in the west could be aided by either reduced wind-driven dune activity or declines in fire frequency. Active fire suppression is rarely attempted in this remote region, however, suggesting wind as the most likely mechanism. The rate of dune movement (creation of ~ 0.14 km^2 $year^{-1}$) is outweighed by the average loss of 1.98 km^2 $year^{-1}$ to forest succession, though these rates are very small relative to the total dune area (~349 km^2). However, we observed an increasing rate of sand dune surface loss between 1985 and 2014, with a net area loss of 1.15 km^2 $year^{-1}$ between 1985 and 2002 and 1.85 km^2 $year^{-1}$ between 1985 and 2014. The total dune area lost to 2014 was 53.76 km^2,

20 percent of the total dune extent in 1985. It is not clear whether this loss pattern is transitory, or whether it may reflect larger climatic changes that favour the stabilization of Athabasca sand dunes.

In summary, our goals were 1) to describe the physical environment of the Athabasca endemics, 2) estimate the distribution of each taxa, and 3) to evaluate the potential impact of dune creation and forest encroachment on these populations. Understanding these factors is critical to evaluating the long-term viability and conservation status of these taxa, particularly the less abundant species. We evidence that the long-term dune stabilization processes are occurring, while currently abundant and widely distributed taxa may be at risk. This is likely due to a historical shift in wind patterns and associated movement of sand to the northeast to the southeast directional range. Reduced dune activity favouring woody vegetation establishment is a significant long-term threat as the preferred habitat of the Athabasca taxa contracts. Development of higher resolution remote sensing strategies to more precisely model the extent of gravel pavements and wet inter-dune slacks using the spectral signatures of exposed gravel and sand surface soil moisture will be important to identify the most critical habitat elements in the landscape. Overall, our current analysis documents the long-term dynamics of this sand dune environment, and how those dynamics relate to the long-term security of the Athabasca sand dune endemic flora.

3.6 Acknowledgements

Fieldwork to collect the ground-truth data was funded by the Saskatchewan Ministry of Tourism, Culture, Parks and Sport, and Environment Canada. Fieldwork was conducted by M. Anderson, G. Argus, D. Guedo, M. Hilderman, S. James, J. Karst, J. Keith, K. Kelly, E. Lamb, G. Longpre, A. Leighton, J. Mischkolz, S. McAdam, C. Neufeld, J. Pepper, J. Smith, A. Tucker, M. Weiss, S. Vinge, B. Wilson, and R. Wright. Funding for the current study was provided by a Natural Sciences and Engineering Research Council (NSERC) of Canada Discovery grant to E. G. Lamb and Department of Plant Sciences support for A.U. Attanayake.

Chapter 4

4 Integration of plant community structure in species distribution modelling: a species co-occurrence based composite approach

4.1 Abstract

Species distribution models (SDM) with remotely sensed (RS) imagery are widely used in ecological studies and conservation planning. The aim of my study was to develop and evaluate mechanisms for improving the prediction accuracies of species distribution models by introducing information on plant community structure, specifically species co-occurrences, into the modelling process. The assessment of plant community structure using image clustering methods found object-based clustered predictors and direct reflectance predictors are following very similar accuracy pattern indicating limited advantages of using high-resolution images. Generally, prediction accuracy is highly correlated with the number of species presences in the landscape. However, the study found for high-frequency species (i.e. >0.5 or present in greater than 50% of plots) that prediction accuracy declines to be as low as the accuracy expected for low-frequency species (i.e. <0.2 or present in less than 20% of plots). The combined influence of both higher false negatives and false positives are contributors to the lower accuracies. The prediction uncertainty of low-frequency species is likely correlated to the unpredictable nature of false negatives. Higher kappa coefficients were obtained species of moderate frequency (present in 20-50% of plots). The study has strong evidence to state that the optimal algorithmic performance is tied to a balanced number of presences and absences in the data. Practically, identification of dominant environmental factors and extension of sampling strategy to include a wide range of such variations likely balance presences and absences for optimal performance. The co-occurrence analysis revealed significant co-occurrence patterns are most common at moderate levels of species occurrence frequencies. The research does not indicate any consistent accuracy increase or decrease between baseline direct reflectance models and composite-SDM framework. Although accuracy changes were marginal with the composite-SDM framework, the method is well capable of influencing associated type 1 and type 2 error rates of the classification.

4.2 Introduction

Species distribution models (SDM) based on remotely sensed (RS) imagery are commonly used in ecological studies and conservation planning. Modelling accuracy with remotely sensed imagery is influenced by the spatial resolution of imagery used, relative plant sizes, and prevalence of the species (Cord and Rödder 2011, Santika 2011, Bradley et al. 2012, Lechner et al. 2012, He et al. 2015, Rocchini et al. 2015). Rarer target species present challenges as small numbers of locations, scattered distribution patterns and differences in realized niche intensify inaccuracies in the prediction process (Santika 2011, Zurell et al. 2016). Compounding these challenges are the non-distinct spectral signatures of some rare grassland plant habitats which are hard to distinguish using air-born or space-born sensors (Zimmermann et al. 2007). Recent SDM studies show that interdisciplinary approaches combining geographical and ecological perspectives into one framework can often positively influence model accuracy (Guisan et al. 2006a, Rocchini et al. 2015, Thorson et al. 2015, Thuiller et al. 2015, Zurell et al. 2016). Here I explore the utility incorporating ecological community information, specifically species co-occurrence patterns, into a composite species distribution modelling framework.

Community assembly processes, species coexistence, and species relative abundances are a result of prevailing abiotic factors and species interactions (Hutchinson 1978, Drake 1990). If such processes are acting to segregate species or to co-locate species, observed co-occurrence patterns are a source of evidence to identify plant species habitat occupancy patterns. Ferrier and Guisan (2006) suggested using an extension of the commonly used individual species distribution models called Stacked Species Distribution Modelling (S-SDM) to predict community composition and species turnover rates. This method initially models individual species as a function of appropriate predictors; individual SDM predicted species distributions are then combined to indirectly estimate species richness patterns. Generally, the stacking process involves individual SDMs being subjected to some form of aggregation, classification, or ordination. The method has been tested with RS predictors, with recent studies suggesting the potential to improve predictions of community composition (Dubuis et al. 2011, Cord et al. 2014a). S-SDM thus uses more widely available co-occurring species records to estimate the probable community composition of the target geographical location (Ferrier and Guisan 2006). Extending the argument, stacking species that are co-occurring at predictable rates with a target species likely

provides a mechanism to improve traditional SDMs. The concept integrates details of ecology of the community into the modelling framework to estimate most probable occurrence extent (Ulrich 2004, Ulrich and J. Gotelli 2007, Thorson et al. 2015, Thuiller et al. 2015, Griffith et al. 2016, Anderson 2017, Ashcroft et al. 2017).

Species co-occurrence patterns are the result of deterministic biotic and abiotic processes (Ulrich 2004, Ulrich et al. 2017). Many community-level metrics have been developed to analyze species co-occurrence patterns (Eg. Gotelli (2000), Ulrich and Gotelli (2010), Ulrich et al. (2017)), however these generally provide little information that can be used to improve SDMs. Pair-wise metrics that examine the co-occurrence of individual pairs of species such as the method proposed by Veech (2013) are readily applicable to SDM. Veech's method identifies non-random co-occurrences (either positive associations or segregations) between pairs of species (Veech 2013), and has generally proved reliable in follow-up testing (Lavender et al. 2019). Pairwise metrics are attractive for incorporation into SDM modelling as they provide straightforward probabilities of association that can readily identify the species that provide relevant distribution information about a focal species.

My main effort is to examine how incorporating pairwise species co-occurrence information into species distribution models can be beneficial to enhance the prediction accuracy for rare and common grassland plant species. Specifically, the study will evaluate 1) the relative importance of ecologically meaningful RS predictors (i.e. object-based analysis) versus pixel-based approaches for modelling grassland plant species, 2) the influence of species distribution frequency on the prediction accuracies by comparing model outcomes between common and rare grassland plant species, 3) the utility of probabilistic co-occurrence analysis approaches to identify sets of co-occurring species suitable for a Composite-SDM framework, and 4) the effectiveness of composite-SDM framework performance with integrated co-occurrence analysis to optimize over and underestimation of the prediction extent for common and rare species respectively.

4.3 Materials and methods

4.3.1 Study area

The study site is Kernen Prairie on the northeast edge of Saskatoon, Saskatchewan, Canada (52°10‟ N, 106° 33‟ W, elevation 510 m, Figure S4.1). The site is a 130 ha remnant fescue mixed prairie containing mosaics of grass and low shrub-dominated communities (Coupland and Brayshaw 1953, Pylypec 1986). Frequent grasses include Plains Rough Fescue (*Festuca altaica* subsp. *hallii*) which grows in association with Wheatgrass (*Elymus lanceolatus*) and Needlegrass (*Hesperostipa curtiseta*). Common forb species include Northern Bedstraw (*Galium boreale)* and Pasture Sage (*Artemisia frigida)*. Parts of the landscape are occupied by low growing shrubs including Western Snowberry *(Symphoricarpos occidentalis)*, Wolf Willow (*Elaeagnus commutata*) and Wild Prairie Rose (*Rosa arkansana)*. The prairie is subject to invasions by Smooth Brome (*Bromus inermis*) and Kentucky Bluegrass (*Poa pratensis*) with Brome patches around the boundaries of the prairie spreading towards the center (Slopek and Lamb 2017). Microtopography, soil water availability, fire, and grazing are major factors structuring the plant community assemblages of the prairie (Coupland and Brayshaw 1953, Romo 2003).

4.3.2 Field data collection

An area of ~10ha of Kernen prairie was excluded from grazing from the beginning of the growing season for this study. Grassland foliage reflectance was measured in July (the period of peak live biomass). Canopy reflectance was measured at a fine-scale using a MicaSense RedEdge multispectral camera (Figure S4.1) capturing five spectral bands including Blue (465-485nm), Green (550-570nm), Red (663-673nm), Red edge (712-722nm) and, Near Infrared (820-860nm) with 47.2° field of view. The camera was flown at 45m altitude on an Unmanned Arial Vehicle (Dragonfly X4P Commander). The height of the flight was maintained to obtain 2.4cm consistent spatial resolution.

Ground-truth species presence-absence data was collected from a field survey of 18 sample points throughout the target landscape. The first sample point was randomly located on the southwest corner of the study site; the rest were systematically located on north or east directions from the first one where each sample point was 50m apart from each other (Figure S4.2). Each

sample point was an 8x8m grid containing 64 1x1m quadrats laid out to the east and south of the sample point (Figure S4.2). I included two extra sample points (S-18 and S-19) off the systematic pattern to ensure surveys of specific plant communities not included in the other plots. The survey recorded all species presences and relative abundances in each 1x1m quadrat. The analysis used 512 quadrats for data analysis.

4.3.3 Raw reflectance vs object-based reflectance derivatives

Building connections between focal species and raw remotely sensed (RS data) is challenging and frequently leads to prediction inaccuracies in SDMs (Cord et al. 2013, Rocchini et al. 2015). This problem is exacerbated for smaller plants relative to large perennial species such as trees that can be measured at coarser pixel scales. In this study employed object-based reflectance derivatives to draw ecologically meaningful environmental and plant community information (geospatial object-based image clustering) (Blaschke et al. 2008). The process of image clustering occupies spatial proximity statistical calculations such that each output cell is based on the surrounding neighbours. The spatial distance was specified by the user and a moving window was used to calculate defined neighbourhood reflectance variation. The focal coefficient of variation was used to summarize the surrounding neighbourhood as it was well recognized object-based image clustering method to identify distance-based reflectance variations in imagery (Jensen 2005, Blaschke et al. 2008). The neighbourhood operation for the target pixel was calculated as a function of all input pixels in defined rectangular proximity. The study used 0.5m and 1m spatial proximity neighbourhoods to identify the spatial dependency in identifying the plant community composition of the landscape. The same process was repeated for all five spectral bands of high-resolution imagery. Please, refer to Figure 4.1 for details of how predictors used in composite species distribution modelling process.

4.3.4 Plant co-occurrence analysis

The species co-occurrence analysis was based on the method proposed by Veech (2013). The method is metric-free, distribution-free and randomization free, and has been found to perform well in comparison to other pairwise co-occurrence methods (Lavender et al. 2019). The analysis was performed using the COOCCUR R package (Griffith et al. 2016). The method requires plant community presence-absence species by site data. The procedure estimates the probability that

two species co-occur at a frequency either significantly less than (negative association), or greater than (positive association) expected compared to a null expectation of random association (Refer to Veech (2013) for theoretical details). The data set was organized with species as rows and sites as columns to comply with the package requirements type= “spp_site” and spp_names = TRUE. The thresh = TRUE criteria was specified to filter species pairs expected to share less than one site. Such species do not have enough co-occurrence information and do not comply with minimum expected frequency of co-occurrence. The decision of random associations with the target species was reached by a heuristic criterion true_rand_classifier = 0.1 that determined random pairs based on less than a 10 percent deviation from their expected number of co-occurring sites. The hypergeometric distribution (prob="hyper") was used to calculate co-occurrence probabilities. Both significant negative and positive associations were used for composite-modelling of the target species as each has the potential to improve respectively predictions of species absence and presence.

4.3.5 Species distribution modelling

The species distribution modelling (SDM) began with observations of species presence/absence. The most critical component of SDM is to identify predictor variables likely to influence or describe the habitat and/or species occupancy. The training process links predictors with the response (species presence/absence) in geographical space to estimate the probability that the species is present or absent in other locations (Figure 4.1). A subset of twelve species was tested with a range of common SDM modelling techniques to identify the preferred method for my study, balancing modelling effectiveness against algorithmic complexity. I tested modern regression techniques that include generalized linear models (GLM), generalized additive models (GAM), and multivariate adaptive regression splines (MARS). Machine learning algorithms tested include classification and regression trees (CART), and artificial neural networks (ANN). The study found GLM performance to be equivalent or better in comparison to other methods (refer to supplementary Figure S4.3), a result consistent with other evaluations of SDM algorithms (Attanayake et al. 2018, Norberg et al. 2019). The model training process and the mathematical formula were produced by BIOMOD2 R package (Wilfried Thuiller et al. 2016, R Core Team 2019).

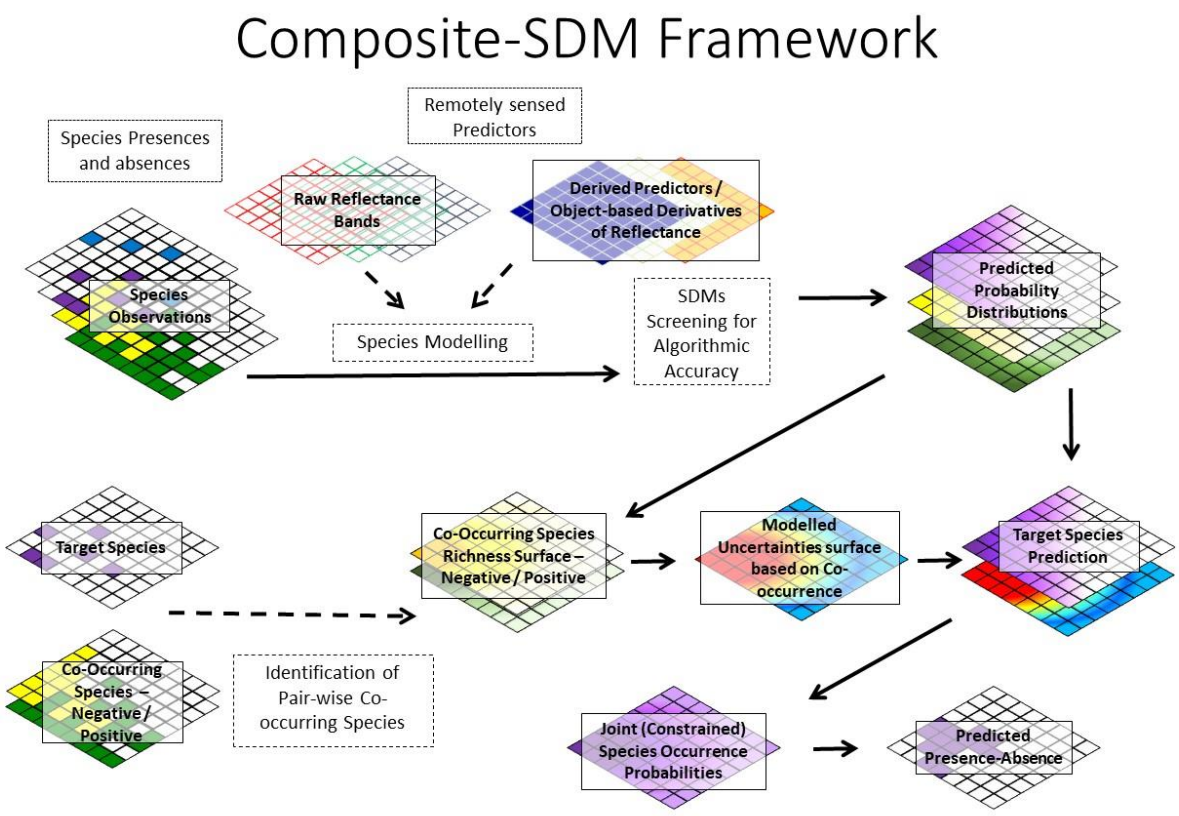


Figure 4.1: Graphical illustration of proposed composite species distribution modelling process. The framework starts with observed species presences and absences. Both direct reflectance and object-based clustered derivatives (focal coefficient of variation) were used as predictors. Species were modelled algorithmically using generalized linear models (GLM). Species co-occurrence analysis was based on the method proposed by Veech (2013)**.** Both significant negative and positive associations were used to constrain the target species predictions. Prediction accuracies were evaluated using 30% of the ground truth data held back from the model training process. I assessed model performance using error of commission (1-User accuracy), error of omission (1-Producer accuracy), overall accuracy, and the kappa statistic.

The study used measures independent of the event in the sample (threshold-independent measures of accuracy) for model selection (Franklin and Miller 2009). The "area under the curve" (AUC) of the receiver-operating characteristic (ROC) plot is well established threshold-independent model selection procedure. The method uses the false positive error rate on the x-axis versus the true positive rate on the y-axis corresponding to each possible value of threshold probability; AUC is calculated by summing the area under the ROC curve where the value is >0.5

(performance better than random). AUC is a reliable measure for model comparisons as AUC is not affected by changes in species prevalence (Manel et al. 2001, Franklin and Miller 2009). The study implemented 100 model iterations for each species and the model with the highest AUC was selected for final species occupancy modelling. Predictive maps of each species were developed in ArcGIS Model Builder.

4.3.6 Composite species distribution modelling

The first step involved identifying pairwise positive, negative, and random associations with the target species (Figure 4.1). Location probabilities of negatively cooccurring space for the target species were calculated by averaging modelled individual species spatial distribution probabilities of the negatively co-occurring species. The same probability averaging process was implemented on the positively co-occurring species to generate positively cooccurring probability space for the target species. Negatively and positively associated continuous geospatial probabilities were converted into a binary surface (presence/absence) using 0.5 as a threshold. The 0.5 threshold was used to ensure an equal chance for a geographic position to be on either side of each surface as a result of averaged location probabilities. Positively associated species locations were given +1 identity and negatively associated positions were given -1 identity for cross-comparison purposes with the target species. The +1 and the -1 location identities were overlaid with predicted target species geographical probability distributions ranging between 0 and 1. For example, overlaying the -1 identity with any location pushes the location value to be within -1 and 0 (0-1=-1 and 1-1=0), while overlaying the +1 identity with any location pushes the location value to be within 1 and 2 (0+1=1 and 1+1=2). The overlaying process was used to constrain primary predictions with the negative and positive geo-spatial identities to generate a target species occupancy pattern in the target plant community based on co-occurrence. When the threshold was applied (Ex – 0.7) on the constrained surface to produce a binary map, always positive co-occurring locations are classified as presences and the negative co-occurrences are classified as absences.

4.3.7 Accuracy assessments

Prediction accuracies were evaluated using 30% of the ground truth data held back from the model training process. The study used 0.7 threshold across all comparisons to facilitate fair

comparisons of binary maps. The 0.7 threshold was chosen based on my previous study (Attanayake et al. 2018). The assessment of model performance was implemented using error of commission (1-User accuracy), error of omission (1-Producer accuracy), overall accuracy, and the kappa statistic. User accuracy is defined as the probability that a predicted class is truly that target class. The probability calculation is dependent on the correctly classified fraction to the total number of locations classified as the target class. Producer accuracy is the probability that a particular class was classified properly in comparison to observations of that class. Cohen's kappa coefficient (κ) measures the inter-rater agreement for qualitative (categorical) items. Kappa coefficients were interpreted using the guidelines outlined by Landis and Koch (1977), where the magnitude of kappa reflects adequate agreement: values < 0 as no agreement; 0.01-0.20 slight; 0.21-0.40 fair; 0.41-0.60 moderate; 0.61-0.80 substantial; 0.81-1.00 almost perfect.

4.4 Results

4.4.1 Study site and plant species occupancy pattern

The survey found 48 species in the study site; Figure 4.2 and Figure 4.3 shows the relative frequency of each species and the modelled probability of occurrences for a sample of species. The most common (*Carex* spp.) was observed in 497 of 512 sample quadrats. *Carex* spp. is predominantly *Carex atherodes*, however other *Carex* species are known from the site and could not be accurately distinguished without flowers (Pylypec 1986). Six taxa were observed in more than 50% of plots: *Carex spp.*, *Poa pratensis*, *Cirsium arvense*, *Sonchus arvensis*, *Galium boreale*, and *Lactuca pulchella*. Thirty-four species were observed in less than 25% of plots. Both the six highly abundant species listed above and some less abundant generalist species (e.g. *Solidago rigida, Rosa arkansana*) were found broadly distributed throughout the landscape. In contrast, most low abundance species (e.g. *Anemone canadensis, Fragaria virginiana, Spiraea alba, Geum macrophyllum, Zizia aptera, Symphyotrichum laeve, Pulsatilla patens,* and *Stellaria longipes*) were only found in specific areas characterized by particular micro-habitat features. For example, *Spiraea alba and Geum macrophyllum* were only recorded inside the Aspen forest patch located on the south-west side of the study site while *Pulsatilla patens* and *Stellaria longipes* occupied more open canopy environments.

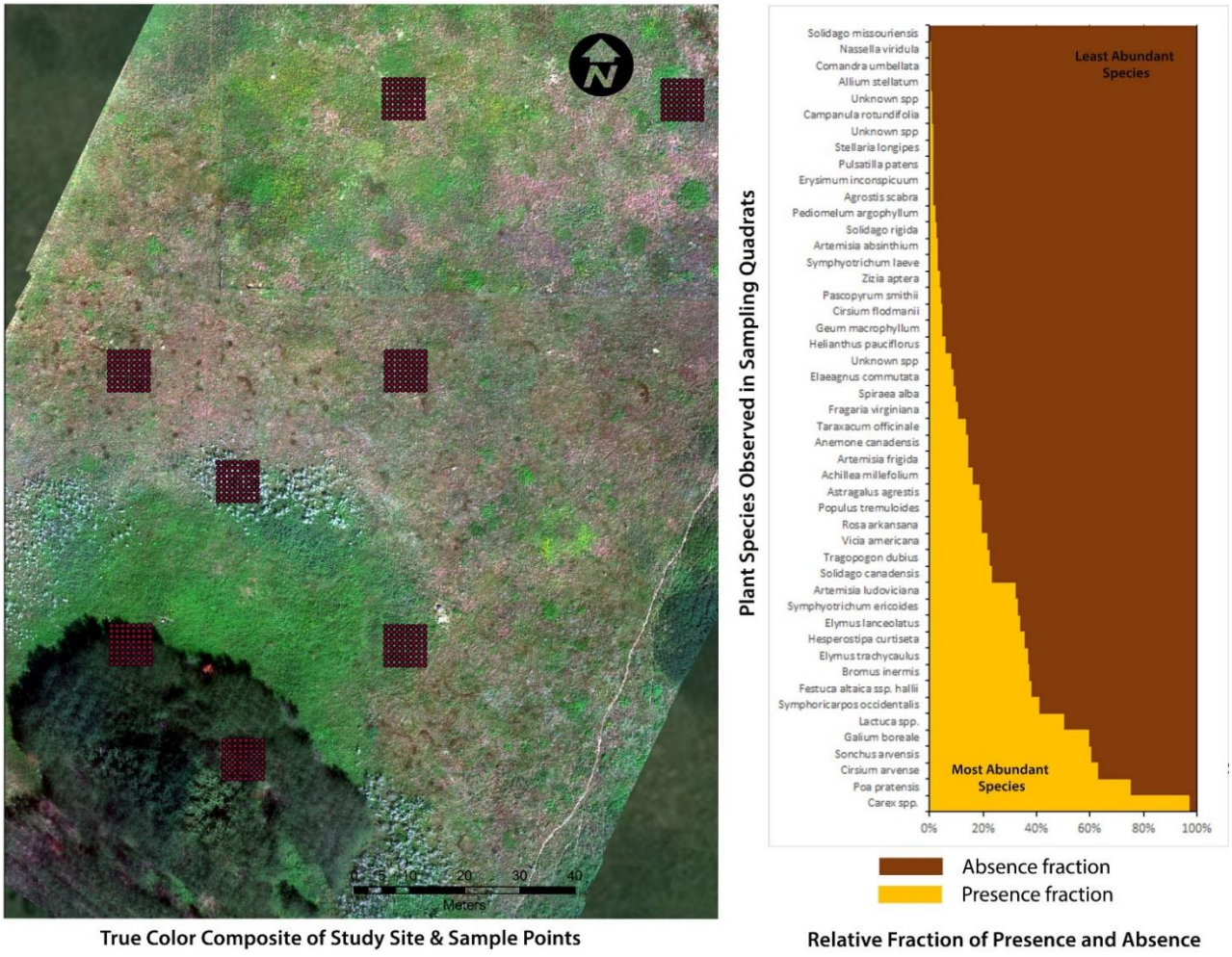


Figure 4.2: a.) True colour composite of the study site with sampling quadrats and b.) Observed species relative presence/absence proportions. The image was obtained from 45m elevation with a multi-spectral sensor mounted on a UAV. The image sensor measured five electromagnetic spectral bands (Red, Green, Blue, Red Edge, and NIR) with 2.24cm spatial resolution. Species presences and relative abundances were measured in 1m quadrats embedded in the 8m by 8m grids at each sample point. The relative fraction of presences and absences for each observed species illustrate the range of species frequency at the study site.

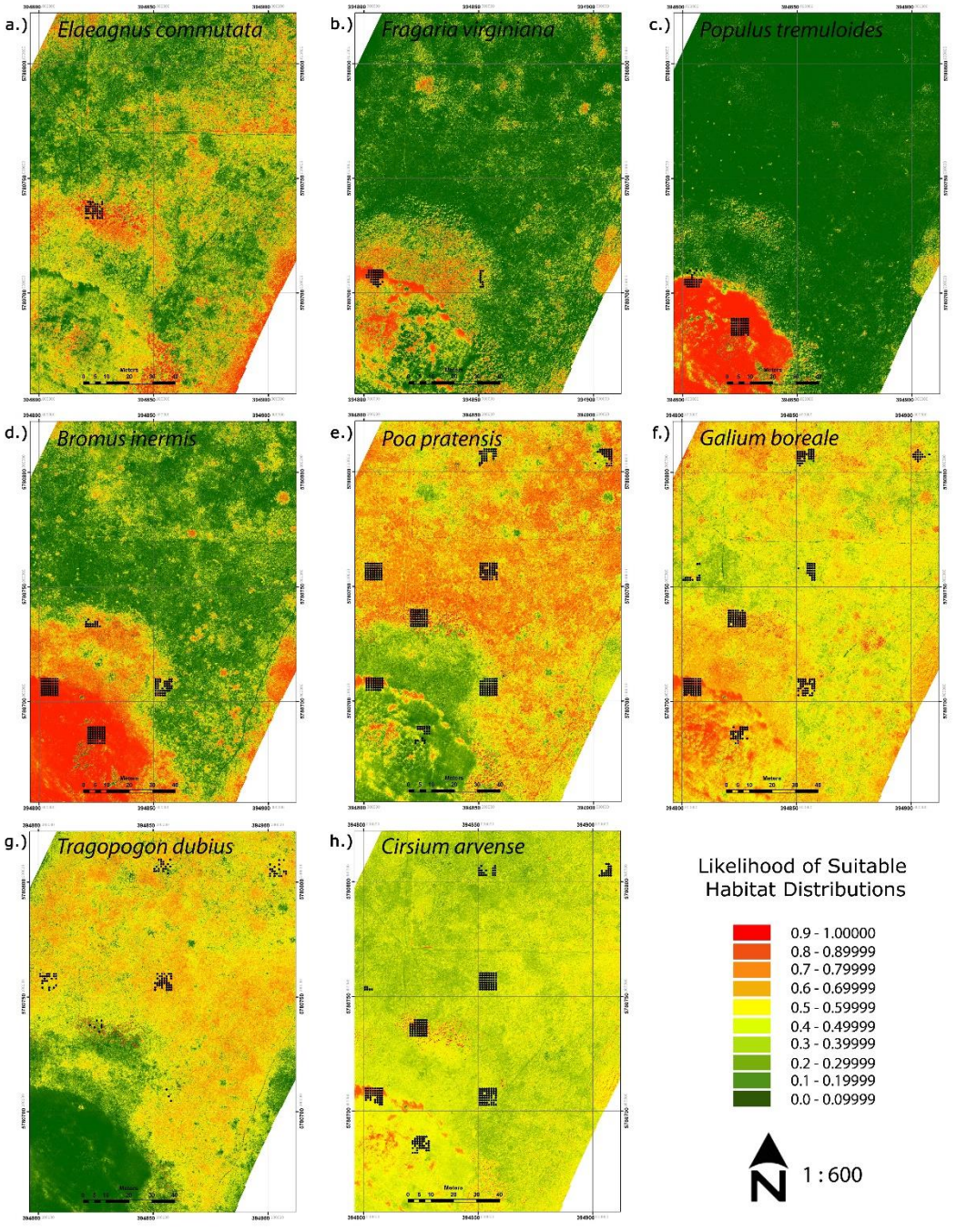


Figure 4.3: Predicted likelihood of suitable habitat distributions for a sample of species observed at Kernen prairie study site. The figure represents the probability of suitable habitat found within a pixel for each of the species. The colour ramp uses ten categories with 0.1 increments. Warm colours indicate a higher probability and the cooler colours indicates a lower probability of finding suitable habitat. Species observed presences are marked with black colour dots.

4.4.2 Species distribution modelling algorithms and predictors

The study used a sample set of species with various habitat characteristics that include *Elaeagnus commutata* (ELACOM), *Populus tremuloides* (POPTRE), *Fragaria virginiana* (FRAVIR), *Bromus inermis* (BRMINE), *Geum macrophyllum* (GEUMAC), *Festuca altaica* (FESALT), *Symphoricarpos occidentalis* (SYMOCC), *Poa pratensis* (POAPRA), *Galium boreale* (GALBOR), *Tragopogon dubius* (TRGDUB), *Solidago rigida* (SOLRIG), *and Carex spp* (CARSPP) to assess performance of several SDM algorithms. In general, the study found above 0.5 AUC values across all sample species for the five different modelling algorithms. The only exception was *Carex spp.* with smaller AUC always below 0.5 which was an indication of uncertain prediction. The assessment generally perceived GLM performance better or very close to other modelling algorithms (Figure S4.3). For example, the mean AUC of *Elaeagnus commutata* across different algorithms were 0.954, 0.778, 0.954, 0.925, 0.938 for ANN, GAM, GLM, MARS, and RF respectively. Thus, GLM was my choice to develop prediction maps for the target species (Figure S4.3). The study further evaluated model performance with two additional variables (plant community height and thickness of litter) as predictors in addition to the reflectance data. These two predictors increased average AUC across most species and substantial AUC increase was observed for *Carex spp.*

Table 4.1: Plant species observed in the study site and pairwise co-occurrence details.

The table includes plant species scientific name, abbreviated term, number of samples ($1m^2$ quadrats) the species was present, the proportion of occupied quadrats (presence fraction), number of positively co-occurring species and, number of negatively co-occurring species. The study site has 512 sample quadrats in total.

Species Abbreviation	**Species Name**	**Species Presences**	**Presence Fraction**	**Positive Co-occurrences**	**Negative Co-occurrences**
ACHI_MIL	*Achillea millefolium*	82	0.160	11	12
AGRT_SCA	*Agrostis scabra*	8	0.016	4	1
ANEM_CAN	*Anemone canadensis*	74	0.145	10	12
ARTE_ABS	*Artemisia absinthium*	15	0.029	5	12
ARTE_FRI	*Artemisia frigida*	75	0.146	9	15
ARTE_LUD	*Artemisia ludoviciana*	165	0.322	17	12
ASTR_AGR	*Astragalus agrestis*	95	0.186	13	11
BROM_INE	*Bromus inermis*	190	0.371	16	16
CARE_SPP	*Carex spp.*	497	0.971	1	3
CIRS_ARV	*Cirsium arvense*	323	0.631	18	10
CIRS_FLO	*Cirsium flodmanii*	24	0.047	4	10
ELAE_COM	*Elaeagnus commutata*	45	0.088	11	12
ELYM_LAN	*Elymus lanceolatus*	174	0.340	9	18
ELYM_TRA	*Elymus trachycaulus*	189	0.369	15	13
ERYS_INC	*Erysimum inconspicuum*	8	0.016	8	2
FEST_ALT	*Festuca altaica ssp. hallii*	195	0.381	20	11
FRAG_VIR	*Fragaria virginiana*	55	0.107	12	14
GALI_BOR	*Galium boreale*	305	0.596	17	10
GEUM_MAC	*Geum macrophyllum*	25	0.049	9	13
HELI_PAU	*Helianthus pauciflorus*	31	0.061	10	9
HESP_CUR	*Hesperostipa curtiseta*	182	0.355	10	20
LACT_SPP	*Lactuca spp.*	257	0.502	13	15
PASC_SMI	*Pascopyrum smithii*	22	0.043	4	3
PEDM_ARG	*Pediomelum argophyllum*	11	0.021	8	2
POA_PRA	*Poa pratensis*	385	0.752	19	7
POPU_TRE	*Populus tremuloides*	99	0.193	10	20
ROSA_ARK	*Rosa arkansana*	101	0.197	13	12
SOLI_CAN	*Solidago canadensis*	119	0.232	16	7
SOLI_RIG	*Solidago rigida*	13	0.025	4	1
SONC_ARV	*Sonchus arvensis*	309	0.604	7	14

SPIR_ALB	*Spiraea alba*	51	0.100	13	15
SYMP_OCC	*Symphoricarpos occidentalis*	210	0.410	20	13
SYMY_ERI	*Symphyotrichum ericoides*	169	0.330	10	16
SYMY_LAE	*Symphyotrichum laeve*	16	0.031	3	3
TARA_OFF	*Taraxacum officinale*	69	0.135	9	9
TRAG_DUB	*Tragopogon dubius*	115	0.225	11	16
UUN1_0825	Unknown spp	41	0.080	7	18
VICI_AME	*Vicia americana*	111	0.217	12	18
ZIZI_APT	*Zizia aptera*	21	0.041	12	11

4.4.3 Plant co-occurrence analysis

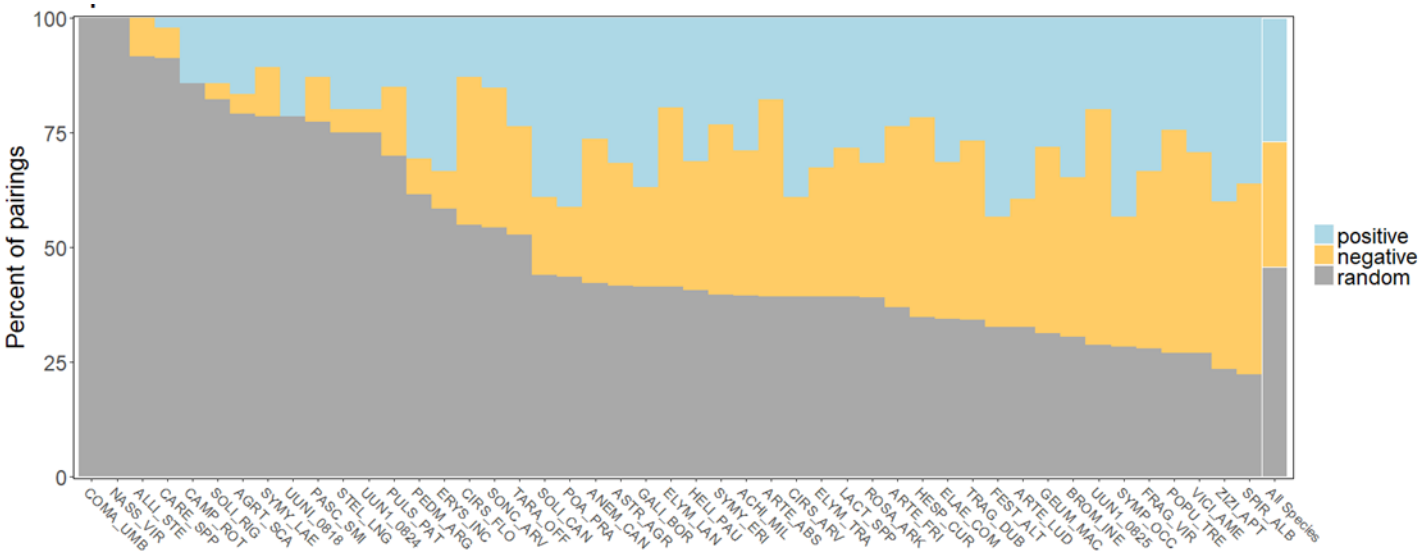


Figure 4.4: Plant species association profile based on co-occurrence analysis. The figure shows the percent of total pairings for each target species divided into pairings with positive, negative and random associations. Species are ordered by decreasing number of random associations. Full plant species names are in Table 4.1.

Numerous positive and negative co-occurrences between species were observed (Table 4.1 and Figure 4.4). Pairwise associations were assessed for a given species pair using probabilities that those species cooccur less- or greater-than what observed. I calculated the percentage of total co-occurring species pairs (positive and negative) for each target species. The association profile ordered from the smallest to the largest is available in Figure 4.4. The analysis shows, *Comandra umbellate, Nassella viridula and Allium stellatum* were not associated with any other species in the plant community (Figure 4.4). The assessment found less than 25% of associations only from 9 species out of 48 in the target community. Larger association patterns were observed from *Spiraea alba, Zizia aptera, Vicia Americana, Populus tremuloides, Fragaria virginiana,* and *Symphoricarpos occidentalis* (Figure 4.4). The study found 28 out of 48 species demonstrate more than fifty percent pairwise association profile with other species. Even with rarest (less than 25% of observed occurrences) species, many co-occurrences were observed. This implies species co-occurrences is not an exception, it is a common scenario. Figure 4.5 shows all pairwise

relationships in the plant community and the sign of the relationships (positive or negative). Figure 4.6 illustrates associations of target species prevalence with the number of positively and negatively cooccurring species. The best fit relationship was quadratic for positively cooccurring number and cubic for negatively co-occurring species, indicating that the highest number of co-occurrences were associated with medium levels of occurrence frequency.

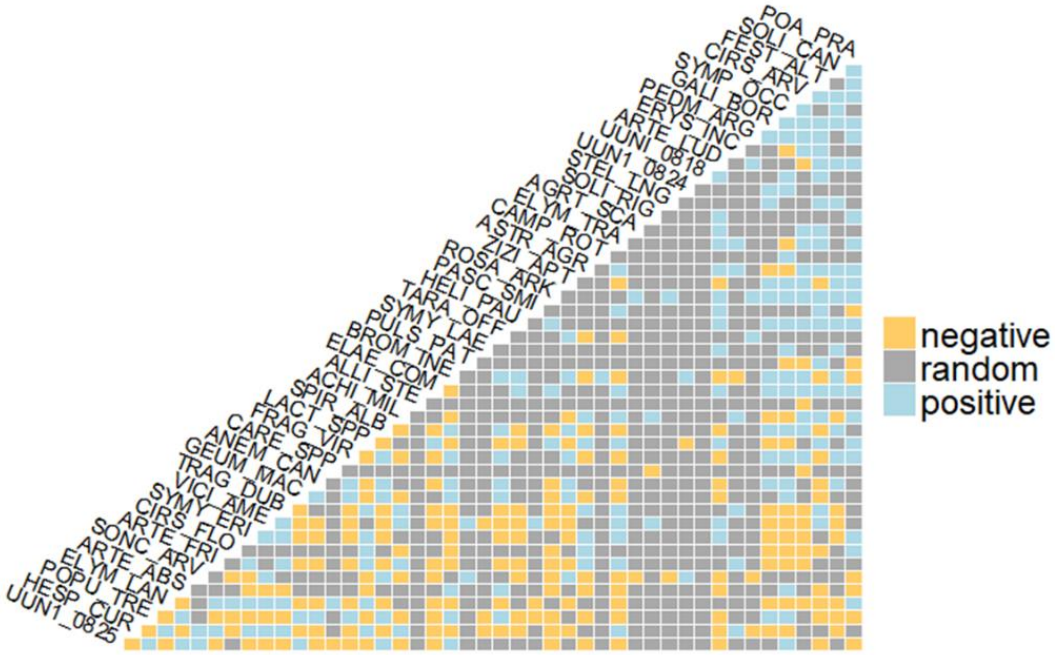


Figure 4.5: Species pairwise associations from probabilistic co-occurrence analysis framework. This analysis determines the degree to which target plant community contains species that are positively, negatively and randomly associated with one another. Species names are organized to represent their relationship with other species in the community.

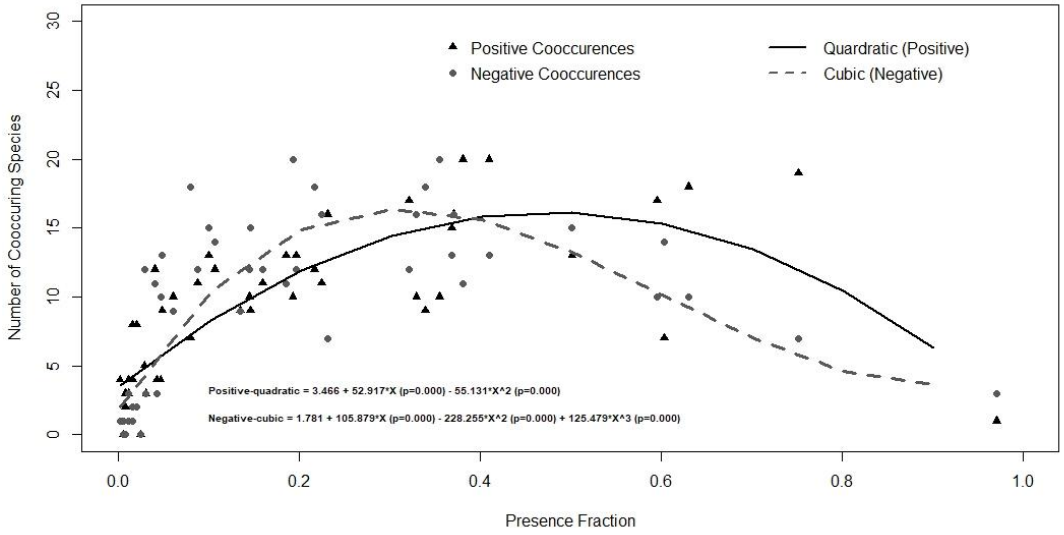


Figure 4.6: Scatter plot and regression of negatively and positively co-occurring species number versus species prevalence. The figure illustrates the density of species positively and negatively co-occur with target species. The regression models were built between the number of co-occurring species and target species observed prevalence. The best fit line was obtained by fitting second and third degree polynomial functions in comparison to a linear regression model.

4.4.4 Species distribution modelling with raw reflectance vs object-based reflectance derivatives

I regressed the species frequencies against overall prediction accuracy with RAW reflectance predictors. The result was a significant decline in overall accuracies with increasing species frequency on the landscape (Figure 4.7). The same gradual decline in overall accuracy with species frequency of distribution was observed with object-based derivative predictors with a spatial neighbourhood of 0.5m and 1m. This similarity implies that the overall accuracy decrease was independent of the neighbourhood influence on reflectance as calculated with object-based derivative predictors (Figure 4.7). Figure 4.8 is a visual representation of species distributions produced by three different predictors for each sample species. I estimated correlations of overall accuracy between direct reflectance, 0.5m, and 1m object-based derivative predictors and the same repeated for kappa coefficient (Figure 4.9). The highest correlations; 0.9 have resulted between

0.5m, and 1m object-based derivative predictors. In general, the study found higher correlation coefficients across all comparisons in both overall accuracies and kappa coefficients.

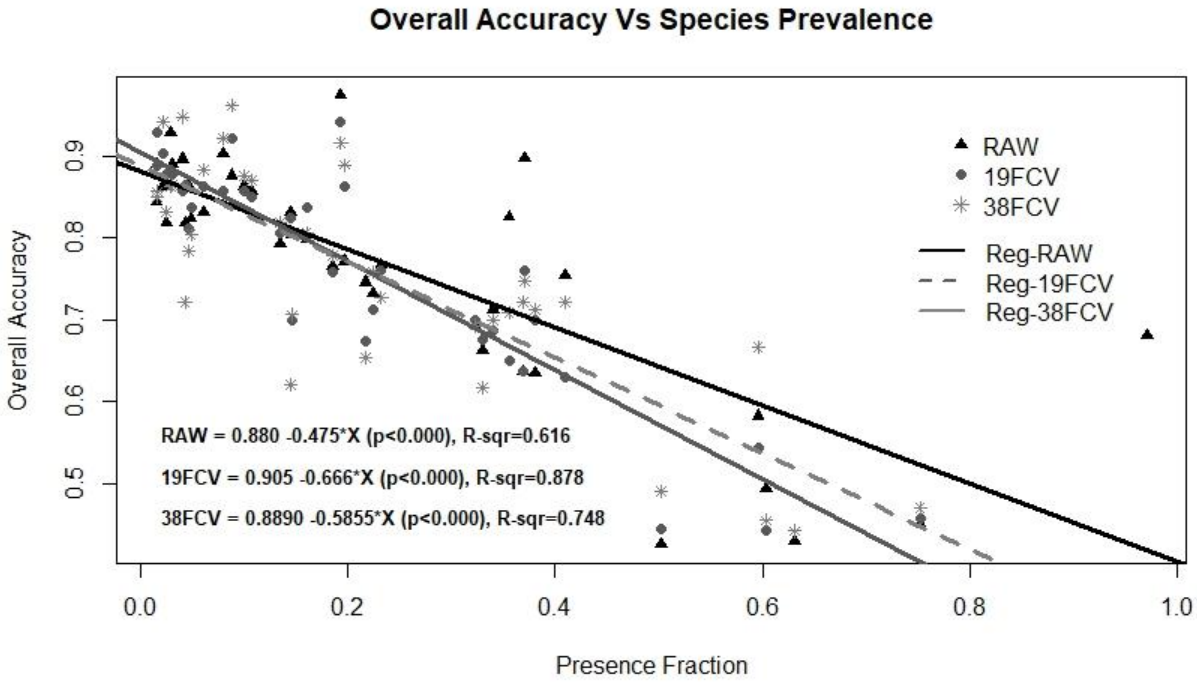


Figure 4.7: Scatter plot and regression of the overall accuracy versus prevalence of species. The scatterplot shows variation in overall model accuracy across three different predictor sets used for species distribution modeling. Baseline models were produced with raw reflectance (RAW) and compared to 0.5m focal coefficient of variation (19FCV) and 1m focal coefficient of variation (38FCV). The regression equation for each predictor category were produced, regressing the overall accuracy against species presence fraction / prevalence.

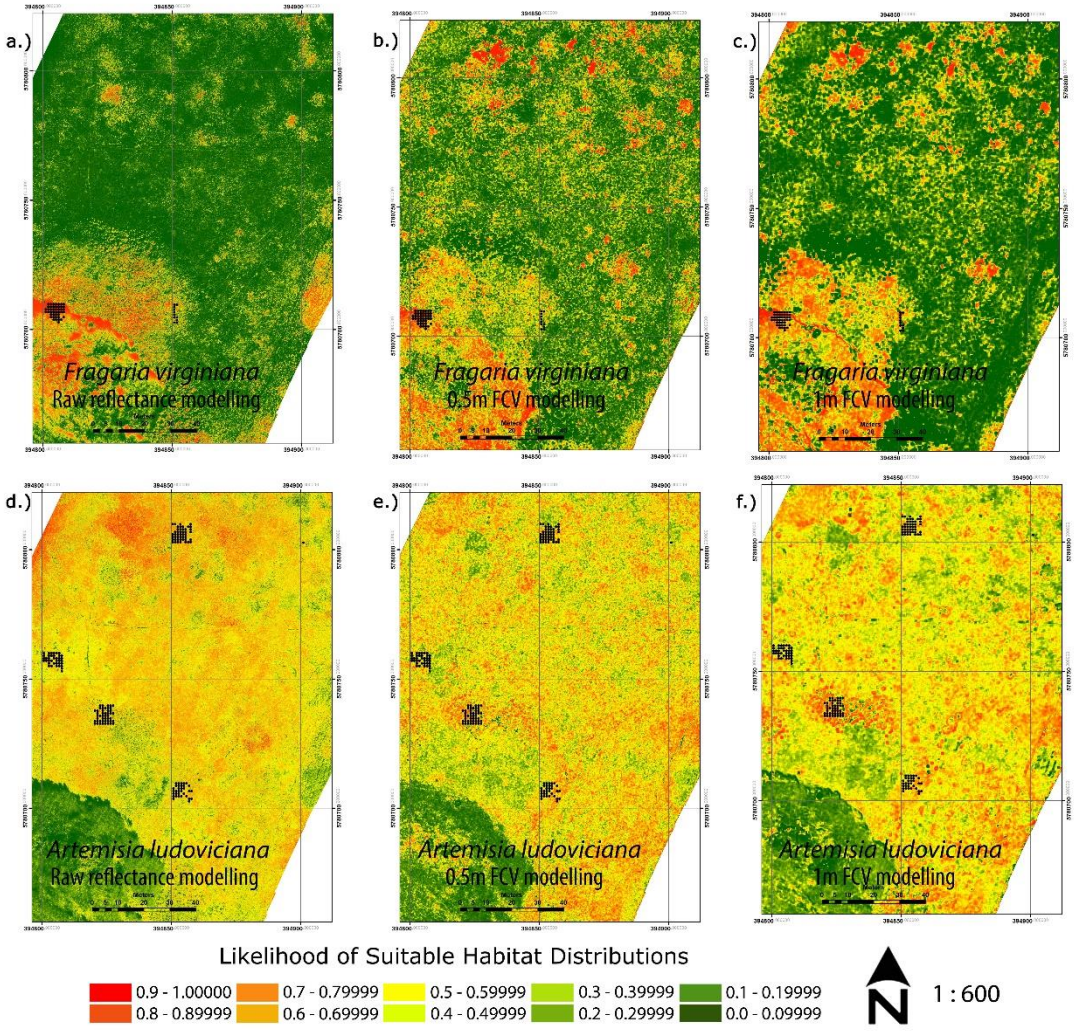


Figure 4.8: Visual presentation of suitable habitat distribution modelling with three different predictors (Raw-reflectance, 0.5m and 1m focal coefficient of variation). The figure represents the probability of suitable habitat found within a pixel for each of the species. The colour ramp uses ten categories with 0.1 increments. Warm colours indicate a higher probability and the cooler colours indicate a lower probability of finding suitable habitat. Two sample species are *Cirsium arvense* and *Fragaria virginiana* and species observed presences are marked with black colour dots.

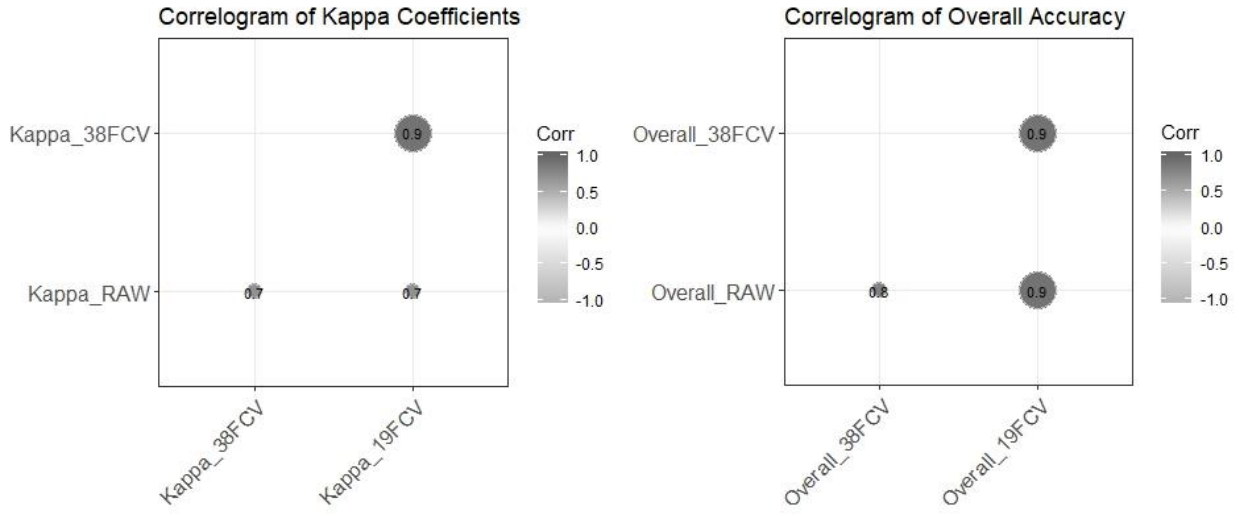


Figure 4.9: Correlogram of kappa coefficients and overall accuracies among predictors. The correlograms represent observed correlations of accuracy measures (kappa coefficient and overall accuracies). The correlations were assessed among each predictor levels used in the study. Baseline models were produced with raw reflectance (RAW) and cross compared to 0.5m focal coefficient of variation (19FCV) and 1m focal coefficient of variation (38FCV).

There were no clear patterns between kappa coefficients of object-based derivative and species frequency (Figure 4.10). I found a significant quadratic relationship of direct reflectance models, indicating that the highest kappa coefficients were likely between frequencies of 0.2 to 0.5. According to Landis and Koch (1977) kappa classification terminology, only *Populus tremuloides (0.91)* has an almost perfect agreement between ground truth and classification. Substantial to moderate agreements have resulted from *Bromus inermis* (0.77), *Hesperostipa curtiseta* (0.6), an unidentified forb (0.52), *Elaeagnus commutata* (0.47), *Symphoricarpos occidentalis* (0.45), and *Fragaria virginiana* (0.4). Fair (0.21-0.40) to slight (0.01-0.2) agreements of classifications were observed for ten and seventeen species respectively. Negative kappa coefficients were observed from six species, indicating no effective agreement between ground truth and classifications. Refer to Figure 4.11-b, Figure 12, and Figure 13 for details.

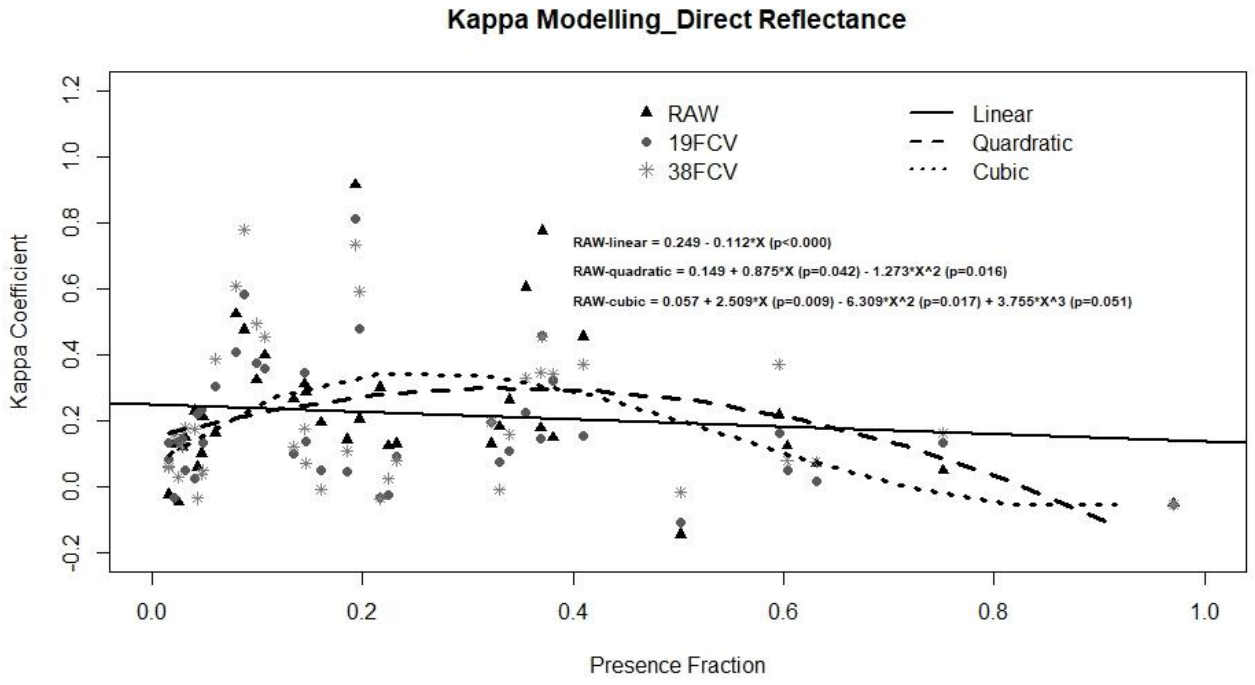


Figure 4.10: Scatter plot and regression of kappa coefficients versus species prevalence (proportion of quadrats occupied). Kappa coefficients are shown for three different predictors used for species distribution modeling. Baseline models were produced with raw reflectance (RAW) and compared to 0.5m focal coefficient of variation (19FCV) and 1m focal coefficient of variation (38FCV). I regressed the kappa coefficients against species prevalence. There were no significant relationships between derived indices (19FCV and 38FCV) and species prevalence. All three regression models were significant for direct reflectance models (RAW). The cubic model was the best out of all three based on AIC (linear -4.279, quadratic -8.668, and cubic -10.969).

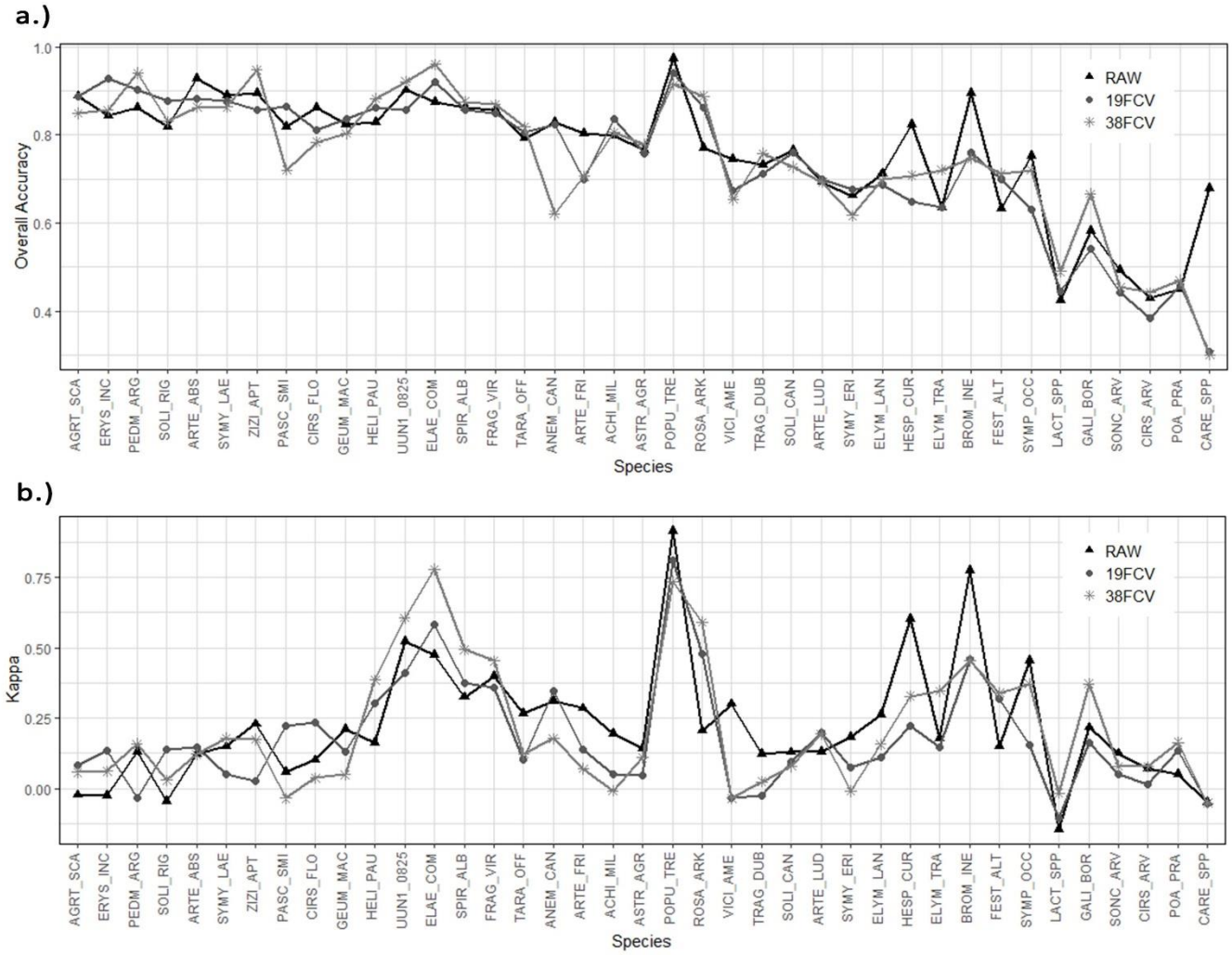


Figure 4.11: a.) Overall accuracy and b.) kappa coefficients for each species observed. The plot represents the overall accuracy and kappa coefficient variations of three different predictor levels used for species distribution modeling. Species are ordered on the x-axis from low prevalence to high prevalence. Baseline models were produced with raw reflectance (RAW) and compared to 0.5m focal coefficient of variation (19FCV) and 1m focal coefficient of variation (38FCV). The accuracy measures are plotted against species ordered from low prevalence to high prevalence.

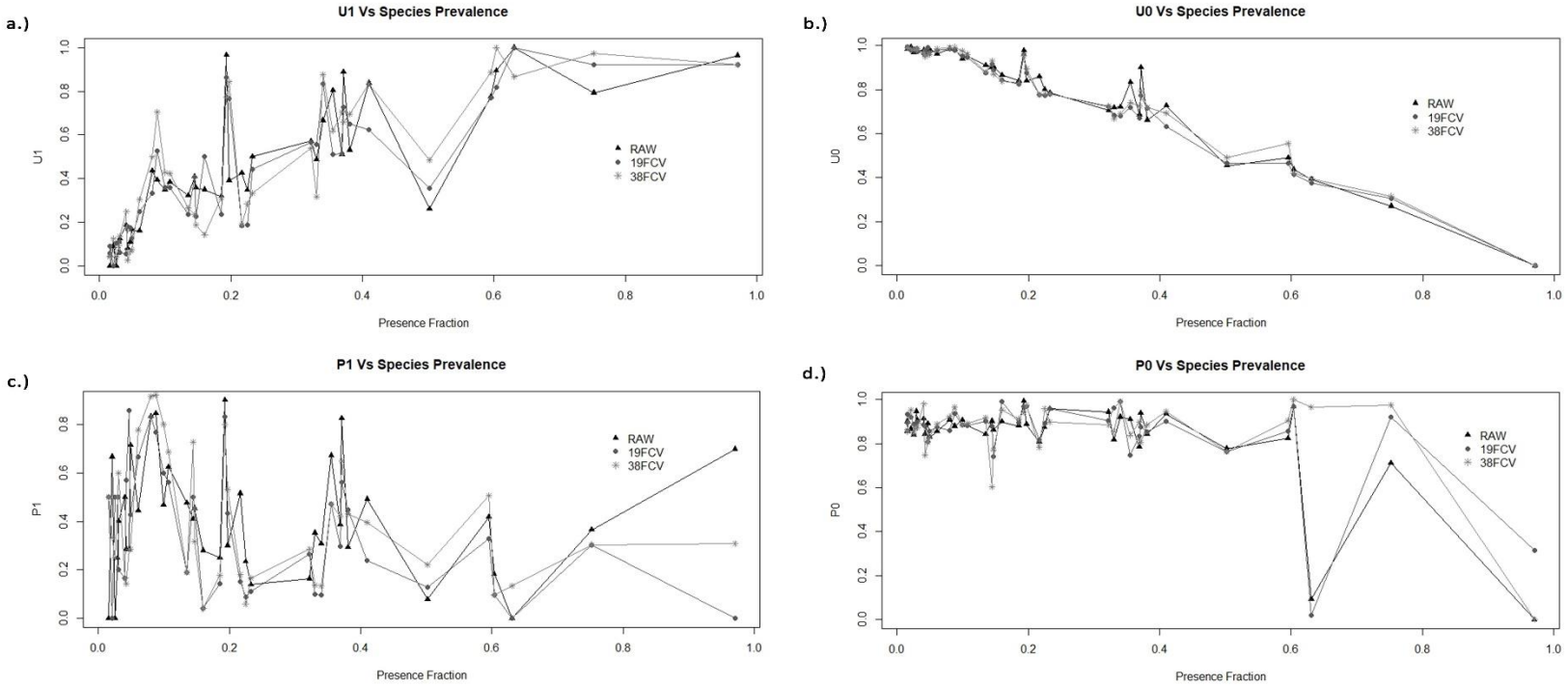


Figure 4.12: Scatter plots of a.) user accuracy of presences (U1), b.) user accuracy of absences (U0), c.) producer accuracy of presences (P1) and, d.) producer accuracy of absences (P0) versus prevalence of species for each predictor level. The plot shows producer and user accuracy variation for each predictor levels used for species distribution modeling. Baseline models were produced with raw reflectance (RAW) and compared to 0.5m focal coefficient of variation (19FCV) and 1m focal coefficient of variation (38FCV). Producer accuracy is the probability that a particular class was classified properly in comparison to observations of that class. Producer accuracy of absence is 1-false negatives and the producer accuracy of presence is 1-false positives. User accuracy is defined as the probability that a predicted class is truly that target class. The probability calculation is dependent on the correctly classified fraction to the total number of locations classified as the target class. The probability is based on the fraction of correctly predicted values to the total number of values predicted to be in a class. The user accuracy of absence is 1-false positives and the user accuracy of presence is 1-false negatives.

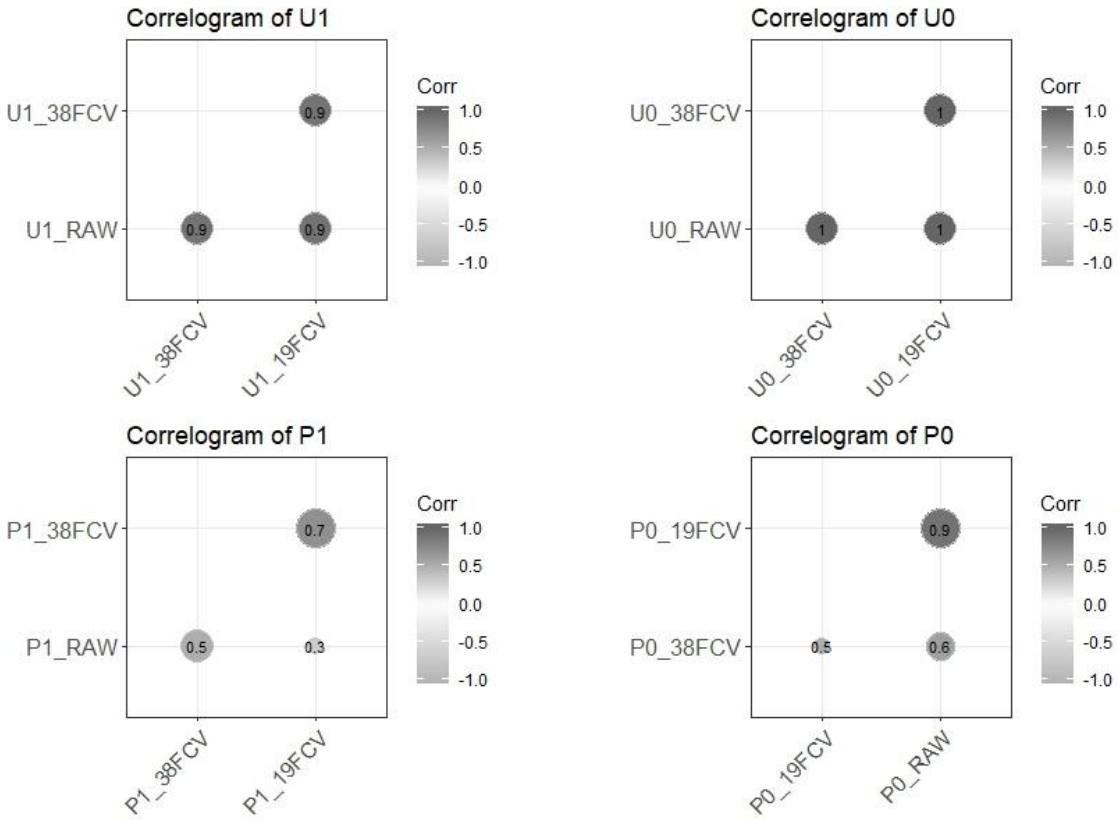


Figure 4.13: Correlogram of user and producer accuracies among predictors. The correlograms show observed correlations of accuracy measures (user accuracy and producer accuracy). The correlations were assessed among each predictor levels used in the study. Baseline models were produced with raw reflectance (RAW) and cross compared to 0.5m focal coefficient of variation (19FCV) and 1m focal coefficient of variation (38FCV). The figure illustrates user accuracy of presences (U1), user accuracy of absences (U0), producer accuracy of presences (P1), and producer accuracy of absences (P0). User accuracy is defined as the probability that a predicted class is truly that target class. The probability calculation is dependent on the correctly classified fraction to the total number of locations classified as the target class. Producer accuracy is the probability that a particular class was classified properly in comparison to observations of that class.

4.4.5 Composite species distribution modelling

The composite-SDM framework was built on the argument that the co-occurrence of a species with other species is not random. If this is true, the observed distribution of a species likely to be a function of co-occurring species. The study compared the composite-SDM framework with direct reflectance model as a baseline to understand the influence of co-occurring species on

the SDM prediction accuracies. Findings state that the result was highly correlated with each other in comparison to direct reflectance models (Figure 4.14). Kappa variations between two series of data were tested for Pearson's product-moment correlation and Spearman's rank correlation rho to identify dependencies. The Pearson's correlation was 0.82 ($t = 8.8669$, $df = 37$, $p<0.001$), in comparison to the Spearman's rank correlation 0.81 ($S = 1826$, $p<0.001$). The result indicates that the composite-SDM framework accuracy closely follows the direct reflectance model accuracy pattern. The overall accuracy has lower correlations due to composite-SDM framework accuracy decline reported from low frequent species (Figure 4.14-a). The Pearson's correlation was 0.62 ($t = 4.8805$, $df = 37$, $p<0.001$) and the Spearman's rank correlation was 0.59 ($S = 4058.6$, $p<0.001$) for overall accuracies.

In general, accuracies are based on several factors; user accuracy (1-commission error on presences or absence) and producer accuracy (1-omission error on presence or absence). User accuracy is calculated based on the probability that a value predicted to be in a certain class really is that class. This calculation is based on the fraction of correctly predicted values to the total predicted to be in that class. The result of user accuracy analysis shows no significant changes in composite-SDM modelling in comparison to direct reflectance models (Figure 4.15 – a and b). The fraction that value in a given class was classified correctly is defined as producer accuracy. The producer accuracies on presences show a general increment (decreased false negatives) (Figure 4.15-c), in contrast, to a decrease in producer accuracies on absence (increased false positives) (Figure 4.15-d). This pattern is very clearly apparent up to mid-levels of species distribution frequencies and the pattern invert thereafter.

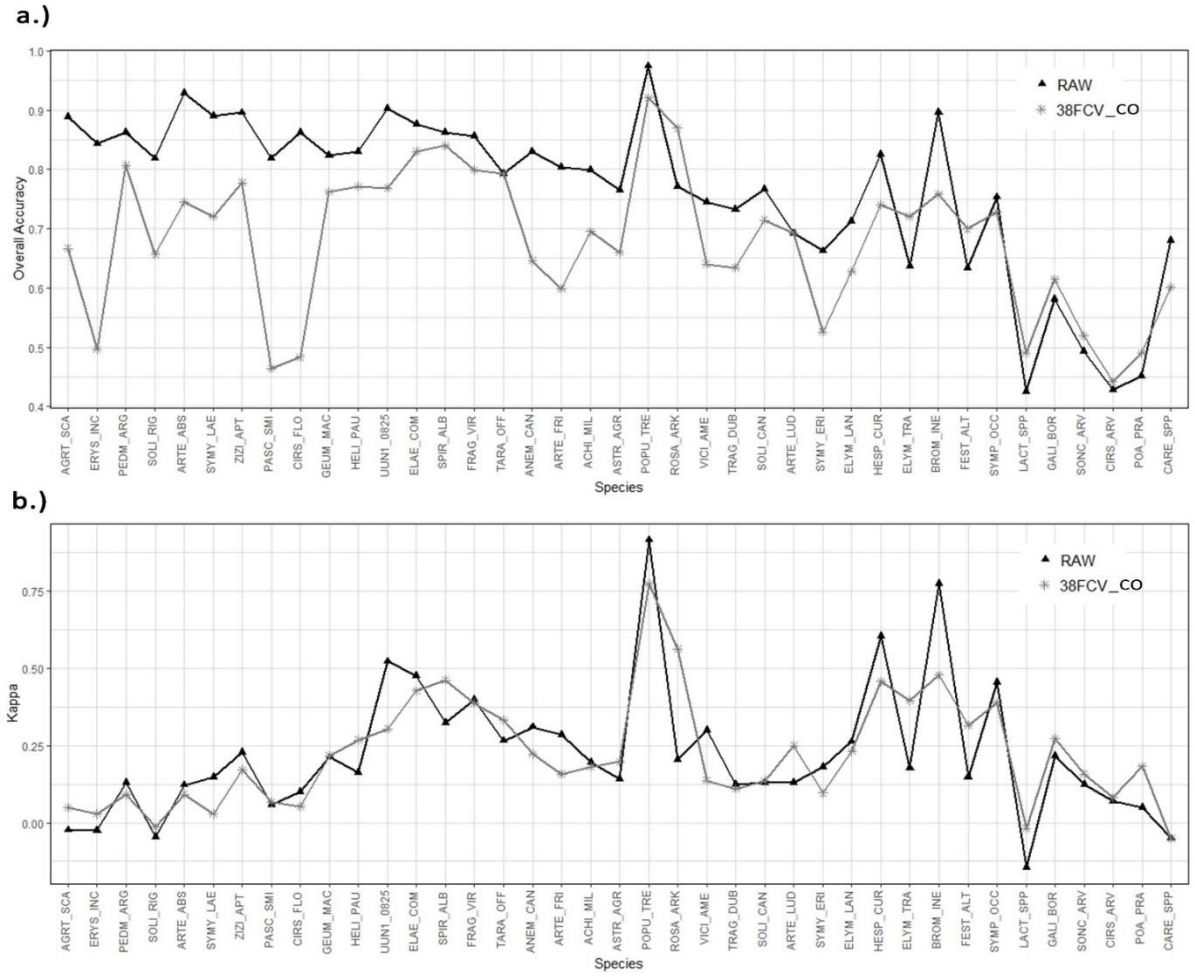


Figure 4.14: a.) Overall accuracy and b.) kappa coefficient variations versus species observed. The plot represents the overall accuracy and kappa coefficient variations of baseline and composite species distribution modeling. Baseline models were produced with raw reflectance (RAW) and cross compared to 1m focal coefficient of variation (38FCV_CO) of composite species distribution models. The accuracy measures are plotted against species ordered from low prevalence to high prevalence.

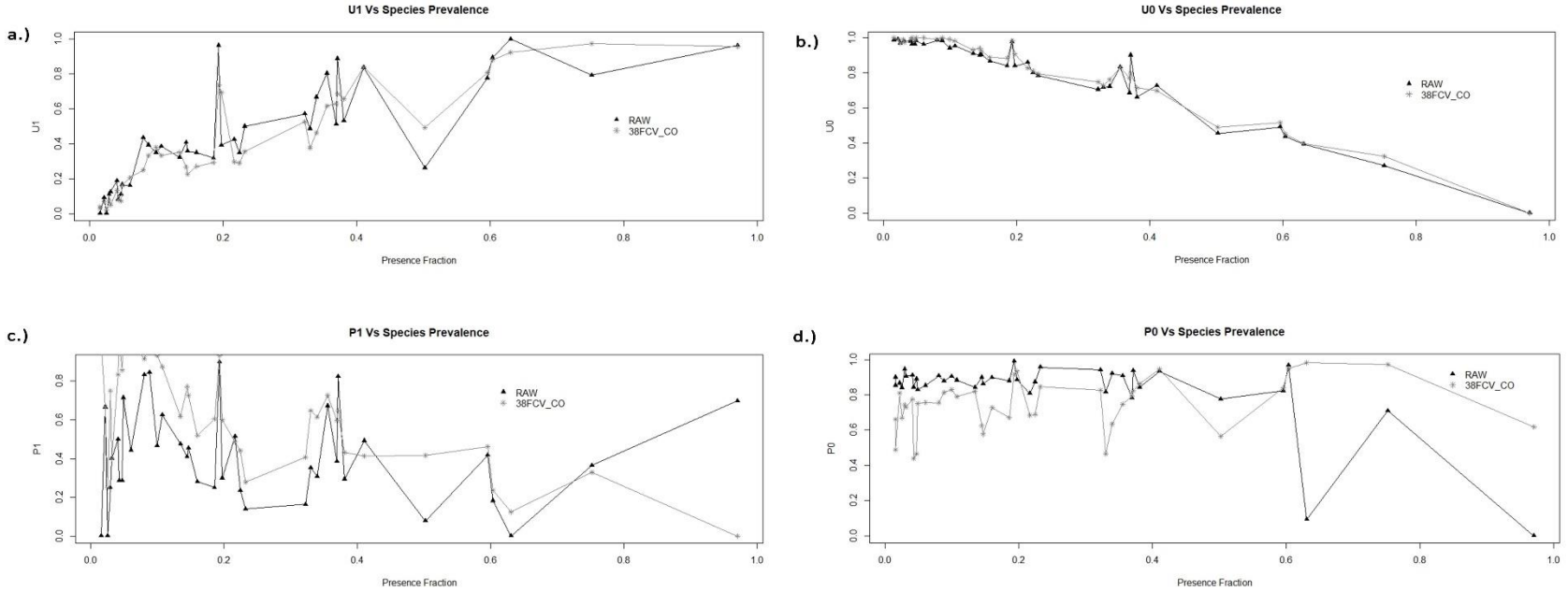


Figure 4.15: Scatter plots of a.) user accuracy of presences (U1), b.) user accuracy of absences (U0), c.) producer accuracy of presences (P1) and, d.) producer accuracy of absences (P0) versus prevalence of species to compare composite species distribution modeling with the baseline. The plot represents the producer and user accuracy variations of baseline and composite species distribution modeling. Baseline models were produced with raw reflectance (RAW) and cross compared to 1m focal coefficient of variation (38FCV_CO) of composite species distribution models. Producer accuracy is the probability that a particular class was classified properly in comparison to observations of that class. The producer accuracy of absence equals to 1-false negatives and the producer accuracy of presence equals to 1-flase positives. User accuracy is defined as the probability that a predicted class is truly that target class. The probability calculation is dependent on the correctly classified fraction to the total number of locations classified as the target class. The probability is based on the fraction of correctly predicted values to the total number of values predicted to be in a class. The user accuracy of absence equals to 1-false positives and the user accuracy of presence equals to 1-flase negatives.

4.5 Discussion

4.5.1 Overview

Species distribution modelling performance is frequently limited by factors including small plant size, small numbers of observations, and scattered distribution patterns due to differences in occupying favourable environmental gradients (Santika 2011, Zurell et al. 2016). The aim of my study was to develop and evaluate mechanisms for improving the prediction accuracies of species distribution models by introducing plant community details (co-occurrences) into the modelling process (Ulrich and J. Gotelli 2007, Ulrich and Gotelli 2010, Veech 2013, Neeson and Mandelik 2014, Veech 2014, Griffith et al. 2016). I first evaluated the relative importance of ecologically meaningful RS predictors (i.e. object-based analysis) versus pixel-based approaches for modelling plant species. Second, I evaluated the influence of species prevalence on model prediction accuracies by comparing model outcomes between common and rare grassland plant species. Third, I assessed the utility of probabilistic co-occurrence analysis approaches to identify sets of co-occurring species suitable for a composite-SDM framework. Finally, I described and tested a composite-SDM framework using the co-occurrence relationships between pairs of species. The study evaluated the influence of classification errors (false positives/type I and false negatives/type II) across all low to high prevalence species surveyed. A similar assessment was used to compare the composite-SDM framework with a baseline reflectance model to better elaborate mechanisms driving model inaccuracies.

4.5.2 Species distribution modelling with raw reflectance vs object-based reflectance derivatives

Image clustering methods were used to test for the effects of the spatial scale of object-based reflectance derivatives on model performance. I tested the role of scale because, while high-resolution images on centimeter or smaller scales such as those from UAV platforms can be obtained, the spatial scales of plant community organization are often 1-2 meters or greater (McNickle et al. 2018). Given the added costs of obtaining high-resolution images, it is important to test for the added value of such images relative to readily available 1-2m scale satellite images. I examined object-based reflectance derivatives with different spatial distances to account for

distance-based spectral variability caused by various species organizations in the plant community. Spatial proximity statistics (coefficient of variation) were calculated such that each location is a representation of surrounding neighbours. In general, I observed similar accuracy between models that used pixel-based direct reflectance predictors and object-based clustered predictors incorporating information from the vicinity.

I noticed both pixel-based and object-based reflectance predictors produced higher accuracies from species such as *Populus tremuloides, Elaeagnus commutata, and Bromus inermis* which are large-bodied prominent indicator species in the plant community (Figure 4.11-b). On the contrary, the majority of both low frequency (<0.2) and high (>0.5) frequency species did not achieve higher accuracies under both circumstances. This indicates limited advantages to using high-resolution images based on similar accuracy patterns observed from both object-based clustered predictors and direct reflectance predictors. Direct reflectance modelling accuracy measures (Overall and kappa accuracies) were highly correlated with those from object-based derivative models with spatial neighbourhoods of 0.5m and 1m. The research findings therefore indirectly imply that there is no significant advantage using 2.45cm resolution in comparison to 0.5m or 1m object-based derivative models. Based on these findings, I argue that 0.5m or 1m image resolutions are sufficient to model species distributions in grassland environments. Biologically these scales make sense as the dominant scale of plant community organization in a Fescue prairie is approximately 1.2m (McNickle et al. 2018), as is the scale of the plant root systems (Lamb et al. 2016). The use of 2.45cm high-resolution images in this context is therefore not necessary and the effort to obtain such images may not be an optimal use of resources, particularly given the greater spatial coverage possible with lower resolution images.

4.5.3 Species distribution modelling accuracy vs species prevalence

Prediction accuracy is generally highly correlated with the frequency of species occurrences in the landscape (Guisan and Thuiller 2005, Guillera-Arroita et al. 2015, Guillera-Arroita 2017). This observation is typically explained based on the assumption that more occurrences allow the algorithm to perform optimally (Thuiller et al. 2009, Guillera-Arroita et al. 2015). In contrast, I found here that for very high-frequency species (i.e. frequency>0.5 or present in greater than 50% of plots) prediction accuracy declines to as low as the accuracy found for low-frequency species (i.e. frequency<0.2 or present in less than 20% of plots) (Figure 4.10). This

opens the question of why kappa-based prediction accuracies are so low for high-frequency species. I suspect that this pattern is arising because most of the high-frequency species in the study system (e.g. *Galium boreale, Sonchus arvensis, Cirsium arvense, Poa pratensis,* and *Carex spp.)* are small-statured habitat generalists. When a species is very widespread, it is likely to be found in a wide variety of ecological contexts, and hence have a wide range of associated species and predictor variable values. Combined with these species being small-statured, and hence unlikely stand out in reflectance data relative to their larger neighbours, the predictor uncertainties associated with widespread species likely act to reduce prediction accuracy. A second explanation is simply that at high prevalence there is insufficient absence data for the algorithm to perform optimally. Insufficient absence data will reduce algorithm ability to identify the clear predictor features of absence.

In general, higher kappa coefficients were obtained for larger-bodied species of moderate frequency (present in 20-50% of plots) such as *Tragopogon dubius, Solidago canadensis, Artemisia ludoviciana, Elymus lanceolatus, Hesperostipa curtiseta, Bromus inermis, Festuca altaica ssp. hallii,* and *Symphoricarpos occidentalis.* These species are either moderately widespread habitat generalists or habitat specialists that cluster in a small subset of the grassland environment. Comparatively bigger plant sizes and densely clustered occurrences dominated by one or few species are likely to create distinct spectral features. Such distinct spectral details and/or dominant environmental preferences are ideal for optimal algorithmic performance (Mateo et al. 2010, Santika 2011). Nevertheless, it is evident that algorithmic performance is generally optimal when there is a balance between presences and absences in the data (Manel et al. 2001, Jiménez-Valverde et al. 2009). Identification of the dominant environmental factors and extension of the sampling area to include areas of poorer habitat for widespread species are strategies likely to produce a balanced number of absences compared to presences in the data.

The biggest challenge was to understand the reasons behind the prediction uncertainties associated with low kappa coefficients for both high and low-frequency species (Figure 4.10). The kappa coefficient was developed based on the concept of inter-rater reliability, a concept which is measured in terms of the observed agreement and expected agreement (Landis and Koch 1977). The highest kappa coefficients occur when high levels of the observed agreement are apparent. The maximum observed agreement is attained when there are no misclassifications (false positives

or false negatives) of presences or absences. The value of the kappa coefficient is thus dependent upon the false positive (type I) and the false negative (type II) rates of classification from both user and producer sides. User accuracy is the probability of predicted class was really is that class and the producer accuracy is the probability that the value in a given class was correct. The study evaluated predictive performance in relation to false positive and false negative rates to explore the factors contributing to increased prediction uncertainty for high and low-frequency species. I found lower false-positive rates for species below a 0.5 frequency of occurrence (Figure 4.12-d). The opposite (higher false positives) was observed for high-frequency species above 0.5 frequency of occurrence (Figure 4.12-d). This implies that the potential for over-prediction due to high false positives is limited for low prevalence species. In contrast, the likelihood of overprediction is high for more frequent species due to the higher rate of false positives. I noticed an apparently random pattern of false-negative variation across the full range of species distribution frequencies (Figure 4.12-c). Overall, the combined influences of uncertainty from higher false negatives and false positives are significant for more frequent species. In comparison, the prediction uncertainty of low-frequency species is tied to the higher number and random nature of the false negatives.

4.5.4 Plant co-occurrence analysis

The pairwise co-occurrence analysis revealed large numbers of both positive and negative co-occurrences among species, with significant co-occurrence patterns most common at moderate levels of species occurrence frequencies. I looked at the percentage of total co-occurring species pairs for each target species and developed an association profile based on positive, negative and random associations (Figure 4.4). Low-frequency species were likely to exhibit totally random associations. Some species like *Campanula rotundifolia, Solidago rigida, Agrostis scabra, Symphyotrichum laeve, Pascopyrum smithii,* and *Stellaria longipes* are associated with around 25% of species out of total observed in the community. Specific habitat preferences are a common feature of species in this group. For example, while *Agrostis scabra* can occur in a wide variety of habitats, I observed dense clusters where the plant community was disturbed along animal trails. Similarly, *Symphyotrichum laeve* prefers the open canopy environment and mesic soils immediately surrounding the aspen forest patch. Other species including *Stellaria longipes, Campanula rotundifolia, and Solidago rigida* are typically found in smaller scattered populations.

I assume such scattered distributions are tied to either preferable environmental conditions or a positive association with other species occurring in that specific area.

I observed the highest proportion of significant co-occurrences with moderately distributed species, more specifically those in the 0.2 to 0.5 frequency of occurrence range (Figure 4.6). This range includes key indicator species, habitat specialists and generalists with clustered distributions. Examples of such species include *Solidago canadensis, Artemisia ludoviciana, Symphyotrichum ericoides, Elymus lanceolatus, Bromus inermis,* and *Symphoricarpos occidentalis*. I found both negative and positive co-occurrences peaked around 0.4 frequency of occurrence. Such strong co-occurrence profiles provide considerable evidence to support the argument that co-occurrence patterns are a promising proxy measure for characterizing the distributions of medium prevalence species. These co-occurrence patterns are similar to those reported by Lavender et al. (2019), who showed that most pairwise co-occurrence tests perform best for medium prevalence species. However, it is not clear why high-frequency species does not show larger association profile.

4.5.5 Composite species distribution modelling

If community assembly mechanisms are a result of deterministic abiotic gradients and species interactions (Hutchinson 1978, Drake 1990), species segregation and co-location can be inferred from observed co-occurrence patterns (Ulrich 2004, Veech 2013, Thuiller et al. 2015, Ulrich et al. 2017). I developed the composite-SDM modelling framework to assess the potential for capturing this information to reduce uncertainty in SDMs. The study included the total plant community observed in the study site to assess how species distribution characteristics based on co-occurrences contribute to the modelling approach proposed. As explained earlier, I used the kappa coefficient to explain model performance as that measure is built on the concept of inter-rater reliability, or a measure of observed agreement and expected agreement (Landis and Koch 1977). The evaluation of model performance is mainly based on classification errors (i.e. false positives and negatives), as all calculation steps for inter-rater reliability rely on correct classification (agreement) and misclassification (disagreement). My results do not indicate any consistent kappa coefficient increase or decrease between the baseline direct reflectance models and the composite-SDM framework. The inconsistent nature of kappa variation was observed across the range of low to high-frequency species. Therefore, while the composite SDM was

overall unsuccessful, there were cases where the community information improved model performance.

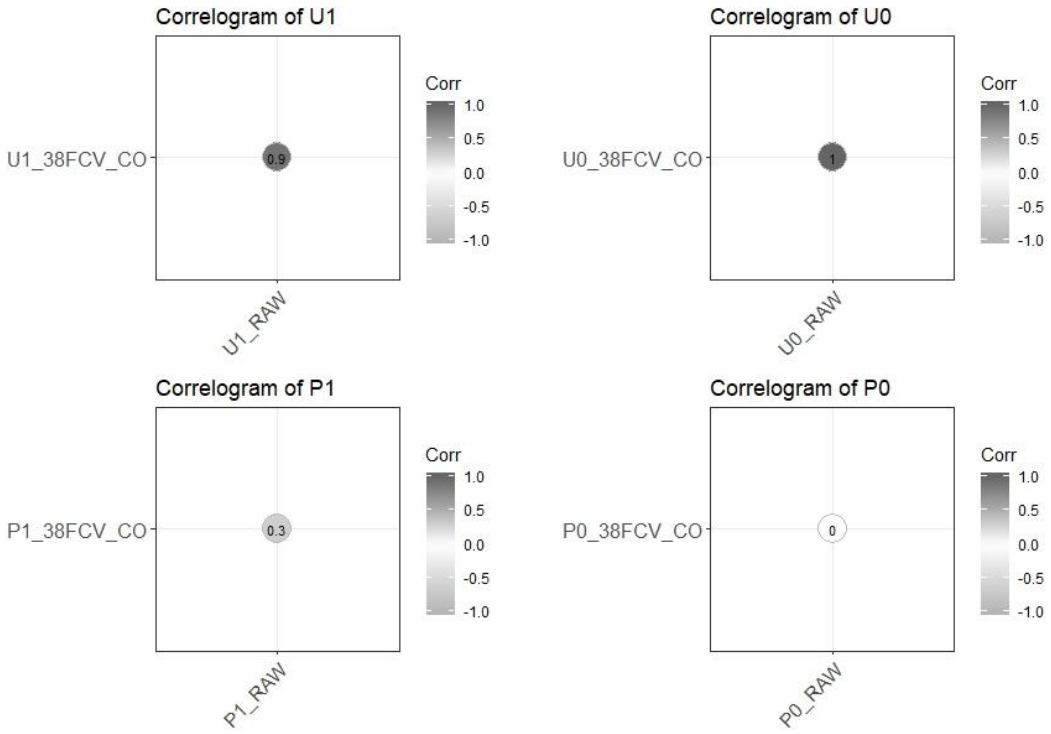


Figure 4.16: Correlograms of user and producer accuracies among predictors to evaluate composite species distribution modeling with the baseline. The correlogram represents observed correlations of accuracy measures (user accuracy and producer accuracy). The correlations were assessed among each predictor levels used in the study. Baseline models were produced with raw reflectance (RAW) and cross compared to 1m focal coefficient of variation (38FCV_CO) of composite species distribution models. The figure illustrates user accuracy of presences (U1), user accuracy of absences (U0), producer accuracy of presences (P1), and producer accuracy of absences (P0). User accuracy is defined as the probability that a predicted class is truly that target class. The probability calculation is dependent on the correctly classified fraction to the total number of locations classified as the target class. Producer accuracy is the probability that a particular class was classified properly in comparison to observations of that class.

To further investigate model performance, I use the probability that the value in a given class was correct (producer accuracy: P1 and P0), rather the probability that a value predicted to be in a certain class was really is that class (User accuracy: U1 and U0). It was apparent that the composite-SDM modelling framework has a great potential to influence both false negatives and false positives consistently (Figure 4.15). I have found that correctly classified probabilities

(presences/P1 and absences/P0) of the composite-SDM framework were not correlated to the baseline direct reflectance model (Figure 4.16). Such a lack of correlation is a result of either increased or decreased classification accuracies in relation to the baseline model. Specifically, the methodology consistently reduced false negatives (1-P1) for species in the distribution frequency region below 0.6 (Figure 4.15 - c). Similarly, I noticed an increased false-positive rate (1-P0) in the same species distribution frequency region (Figure 4.15 - d). More abundant species (greater than 0.6 distribution frequency) modelled with the composite-SDM framework showed higher false negatives (1-P1) and lower false positive (1-P0). Refer to Figures 4.15 c and d for graphical illustrations of the false negative (1-P1) and false positive (1-P0) rates in relation to species frequency.

Such mutual dependencies of low false negatives with elevated false-positives or vice versa is the main reason why the research did not observe an increased overall inter-rater agreement in the composite-SDM framework relative to baseline models. Usually, the kappa coefficient goes up when there are no misclassifications from both presences and absences (Landis and Koch 1977). In the proposed methodology, the decreases in false negatives (increased correctly classified presence), increases false positives (decreased correctly classified absence). Otherwise, it goes in totally opposite directions producing high false negatives and lowers false-positives that leads to marginal kappa results. The classification accuracy of predictions is mainly based on the binary map produced using a threshold chosen. Usually, an increase or decrease in the threshold can influence the accuracy results and finding an optimal solution is always context-specific. The consistent nature of the accuracy changes along species distribution frequency gradients suggests that an examination of different thresholds might be another path to optimize prediction accuracies within the composite framework. I separately averaged the predicted probability surfaces of positive and negative co-occurring species to produce a co-occurring geographical distribution of negative and positive locations for the target species. It may be possible to replace the averaging process of negative and positive co-occurring geo-spatial probabilities with alternative algorithms such as choosing the best principle components. Such algorithmic alternatives are likely to optimize geospatial co-occurrence probabilities for the target species and stabilize inaccuracies in the composite-SDM process.

Chapter 5

5 General discussion

5.1 Overview

Species distribution modelling (SDM) performance is frequently limited by factors including small plant size, small numbers of observations, and scattered distribution patterns (Santika 2011, Zurell et al. 2016). The focus of my thesis was to develop and evaluate alternative SDM methodologies to deal with such challenges. I used two data sets, a record of endemic species occurrences across 350km^2 from the Athabasca sand dunes in northern Saskatchewan (Lamb and Guedo 2012), and small-scale plant occurrence and UAV imagery from Kernen Prairie, a remnant Fescue prairie in Saskatoon, Saskatchewan (Pylypec 1986). Specifically, the Athabasca endemic data were used to estimate the population size and habitat extent for each endemic species, and to describe the dune environment by evaluating dune morphologies, long-term dune spatio-temporal variations, and rates of woody vegetation encroachment and dune stabilization to evaluate an important potential threat to the Athabasca endemic flora. This work is one of the first to use SDM methods to assess plant species population sizes for the purpose of conservation status assessment (COSEWIC 2018).

I evaluated SDM methodologies, first by assessing five different modelling algorithms including modern regression and machine learning techniques to understand how species distribution characteristics influence prediction accuracies using the same data. The primary aim of my work was a series of studies to develop and evaluate mechanisms for improving the prediction accuracies of species distribution models by introducing plant community details (co-occurrences) into a composite modelling process. My attempt was structured around evaluating the relative importance of object-based analysis versus pixel-based approaches to account for plant community details, the influence of species prevalence on model prediction accuracies by comparing model outcomes between common and rare grassland plant species, the utility of probabilistic co-occurrence analysis to identify sets of co-occurring species, and test a composite-SDM framework using the co-occurrence relationships between pairs of species. Understanding the influence of classification errors (false positives and false negatives) of the composite-SDM

framework in relation to different prevalence patterns of species was a crucial element to better examine mechanisms involved with inaccuracies.

In the following sections, I broadly discuss and integrate the overall findings of my thesis. First, I consider SDM performance and algorithm selection in relation to species characteristics and distribution patterns. Second, I evaluate the importance of Remotely Sensed (RS) direct reflectance predictors with object-based reflectance derivatives to point out the influence of spatial variability on the prediction accuracies. Third, I consider species co-occurrences and how to integrate such interspecies relationships to model other species. This is a composite approach to SDM integrating plant community details rather individual species modelling with pre-defined predictors. Finally, I discuss technical limitations, application challenges, and new directions.

5.2 Influence of species characteristics and distribution pattern on the algorithm performances

In the study, I used five different modelling algorithms including both modern regression and machine learning techniques: generalized linear models (GLM), generalized additive models (GAM), multivariate adaptive regression splines (MARS), classification and regression trees (CART), and Artificial neural networks (ANN). The effectiveness of the five algorithms was evaluated in light of the distributions of the species and the spatial resolutions of predictors. The tests clearly show that each method tested was capable of handling low to high frequent species. This conclusion was reached based on the robust performance of all algorithms (>0.5 AUC). Strong GLM performance was observed irrespective of the species distribution pattern and from both study sites. Strong linear relationships between species habitat occupancy and unique features of the reflectance bands that distinguish presence from absence are likely the reason driving such strong predictive performance. However, it should be emphasized that the optimum GLM performance was observed at the medium levels of species occurrence frequencies (i.e. species present in 20% to 50% of plots). Highly uncertain GLM performance was obtained for both low and high-frequency species. Similar results were apparent from other algorithms tested in the study. Therefore, care should be taken at the selection stage as the complexity of algorithm increases from GLM, GAM, MARS, CART, to ANN. Usually, computational intensity and the time taken to produce final results are highly associated with the number of data points in the

dataset and algorithmic complexities. Given these constraints, I suggest the use of GLM whenever possible, as GLM is computationally efficient, but does not compromise accuracy for simplicity.

In general, higher accuracies were obtained for larger-bodied species of moderate frequency (present in 20-50% of plots). These species are either moderately widespread habitat generalists or habitat specialists that cluster in a small subset within the environment. Comparatively bigger plant sizes and densely clustered occurrences dominated by one or few species are likely to create unique spatial patterns that contribute to distinct spectral features. Another reason for higher accuracy is that habitat specialists often rely on spatially restricted and spectrally distinct environmental conditions (Guisan et al. 2006a, Franklin and Miller 2009). Such distinct spectral details of dominant environmental preferences are ideal for optimal algorithmic performance (Mateo et al. 2010, Santika 2011). Identification of the dominant environmental factors and extension of the sampling area to include wider gradient of assumed determinants of species existence likely to stabilize the imbalances of data.

I studied various species distribution patterns of a common and rare sand dune and grassland plant species. A consistent observation in my work was that SDM prediction accuracy is highly correlated with the frequency of species presences in the landscape. The importance of presence frequency is extensively discussed in the literature (Guisan and Thuiller 2005, Guillera-Arroita et al. 2015, Guillera-Arroita 2017), and is typically associated with the assumption that algorithm performance improves the higher the number of species presences and the stronger the environmental determinants as predictors (Thuiller et al. 2009, Guillera-Arroita et al. 2015). The balance between presences and absences is assumed to be naturally reached when species presence frequency increases relative to the total sample size. Such a balance between both presences and absences in the sample produce an ideal setting for optimum performance. Critical in study design is a sampling design that includes all plausible habitats of the target species to avoid small numbers of presences whenever possible.

While the implicit assumption in the literature is that more presences always improves performance (Guisan et al. 2006b, Elith and Graham 2009, Franklin and Miller 2009), for very high-frequency species (i.e. frequency>0.5 or present in greater than 50% of plots) prediction accuracy declines to as low as the accuracy expected for a low frequency species (i.e. frequency <0.2 or present in less than 20% of plots). I propose that this pattern is arising because most of the

high-frequency species in my grassland study system are small-statured habitat generalists. Such species generally contribute little to the electromagnetic reflectance captured by the imagery used to model the target species. This is where widespread habitat specialist species or larger-bodied species have the advantage to attain higher predictive performance. When a species is very widespread, it is likely to be found in a wide variety of ecological contexts, and hence a wide range of geographical predictor variabilities. Combined with high prevalence species often being small-statured species and hence unlikely to stand out in reflectance data relative to their larger neighbours, these uncertainties likely act to reduce prediction accuracy. If minimizing predictor variability going to be a solution, geographical partitioning of predictors is likely to minimize associated geospatial predictor inconsistencies with presences. The modelling process should partition the study site initially using prominent vegetation or similar criteria to generate relatively homogeneous microclimatic or micro-community clusters. Unsupervised classification of the study site is another solution if there are no clear criteria to use for clustering. The species should be modelled separately in each geographical cluster such that the species is effectively a specialist species in that particular habitat.

Insufficient absence data for the algorithm to perform optimally in predicting absences is another issue associated with widely spread species. This is another instance where the data loses the balance between presences and absences. Most algorithms tested in the study proved to work more precisely when there is a balance. Just like the problems that arise with insufficient presence data, insufficient absence data will reduce the algorithm's ability to identify clear absence predictor features. This will minimize the ability to predict nun occurrences relative to occurrences and will increase both false positives and negatives, contributing to lower kappa coefficients. A common solution to deal with absences is the inclusion of pseudo absences into the modelling framework to bring balance to the data (Engler et al. 2004, Lobo and Tognelli 2011, Barbet-Massin et al. 2012). Extreme care should be taken to choose locations for pseudo absences that avoid further confusion with associated predictors. This can be achieved by applying unsupervised classification to the image and segment geographical localities which are similar to observed absences. Such methods will make sure pseudo absences are located in similar reflectance regions of the image where real absences were observed.

The common practice of transforming predicted location probabilities into presences or absences requires a specific threshold to be chosen for conversion (Jiménez-Valverde and Lobo 2007, Franklin and Miller 2009, Liu et al. 2013). Predictive performance of any modelling effort is widely influenced by the threshold used to generate binary maps (predicted presences and absences). The binary map produced is the input used for confusion matrix-based classification accuracy criteria. I looked at how habitat predictions can be influenced by different threshold probabilities from the moderate to the highest using the Athabasca endemic species. In general, increasing threshold probability had a decreasing trend of estimated presences irrespective of species occurrence frequency. As often expected, estimated presences were low when there is a smaller observed habitat extent compared to higher estimated presences for common species that had a large observed habitat extent. This trend was observed up to the 0.6 threshold level, however the pattern of estimated relative abundances drastically deviated from the observed pattern beyond the threshold of 0.6. These results therefore show that picking optimum levels for the threshold is of utmost importance to reach correct predictions. The analysis highlights that choosing the highest threshold is not ideal as higher thresholds deliver inconsistent predictions compared to observed patterns of occurrence frequency irrespective of species prevalence. I recommend to carefully evaluate and select the most appropriate threshold for each situation; the only exception is cases where the study is comparative and the same threshold must be applied across all species. There are many subjective and objective approaches to determining the threshold that has been developed. The fixed threshold approach (i.e. using 0.5 as the threshold) is a subjective approach (Liu et al. 2013). Objective approaches include Kappa maximization, overall prediction success, average probability, mid-point probability, and ROC plot-based methods (Liu et al. 2005, Freeman and Moisen 2008, Bean et al. 2012). Use of such criteria to determine the best threshold is highly recommended for any practical application as species with challenging characteristics (low prevalence) usually sensitive to the choice of threshold. The selection of the criteria is context-specific and I always prefer to provide a continuous probability surface allowing users to select threshold choice based on the acceptable rate of prediction accuracy and precision.

5.3 Predictors - direct reflectance and object-based derivatives

Use of per-pixel information usually does not account for neighbourhood characteristics. Pixel-based information provides only surface details of the target defined by the sensor resolution

(i.e. spatial, spectral and radiometric). The most significant constraint with higher resolution imagery is higher within class classification variability due to the fact that each object is composed of many pixels (Blaschke et al. 2008, Weng 2011). This situation is significant when pixel density increases with high-resolution imaging. My study used image clustering methods to test for the effects of the spatial scale of object-based reflectance derivatives on model performance. In general, I found similar accuracy performances between models that used pixel-based direct reflectance predictors and object-based clustered predictors incorporating information from the vicinity. Furthermore, both methods produced higher accuracies from species which are large-bodied prominent indicator species in the plant community. On the contrary, the majority of both low frequency (<0.2) and high (>0.5) frequency species did not achieve higher accuracies under either circumstance. It is apparent from the study that the distance-based reflectance standardization using coefficients of variation, although producing visually appealing clustering results, made few real contributions to improve accuracies. Segmentations with contextual details of the target such as cluster size, cluster shape, relative location, boundary variability, and topological relationships are all likely positive influences to reach better clustering results (Blaschke et al. 2014). This is called Geographical-Object-Based Image Analysis composed of many interesting dimensions to get groups of pixels into the context that is crucial for understanding ecological constituents of plant community structural variations. Inclusion of such details in the clustering algorithm compared to distance-based CV is time and computationally intensive.

My results indicate limited advantages of using high-resolution images based on the similar accuracy patterns observed from both object-based clustered predictors and direct reflectance predictors. The clustering process standardized each pixel reflectance relative to the pre-defined neighbourhood to integrate distance-based reflectance variations of the plant community. The process assumes that standardizing pixels based on surrounding reflectance variability will minimize the within object variability associated with high-resolution images. The result does not provide compelling evidence that there is a variability difference between the 1m scale and the direct 2.4cm scale. If the environment does not hold useful information at higher resolution scales, there is no practical reason to use a 2.4cm scale. In similar observations, the spatial scales of plant community organization have been found to be often 1-2 meters or greater (McNickle et al. 2018). Biologically, the 1m scale observed in my study makes sense as the dominant spatial scales of

plant community organization in Fescue prairie is approximately 1.2m (McNickle et al. 2018), as is the scale of plant rooting systems (Lamb et al. 2016). The use of 2.45cm high-resolution images at this context is therefore, not necessary and the effort to obtain such images is not an optimal use of resources, particularly given the greater spatial coverage possible with lower resolution images. Similar modelling procedure with low spatial resolution images (0.5, 1,2,5, and 10m) are highly recommended to further validate research findings and to identify the optimal spatial scale for prediction.

5.4 Plant co-occurrence analysis and composite species distribution modelling

The pairwise co-occurrence analysis was the base for the composite-SDM framework. The co-occurrence analysis revealed large numbers of both positive and negative co-occurrences among species, with significant co-occurrence patterns most common at moderate frequencies (0.2 to 0.5) and peaking at around 0.4 frequency of occurrence. Most high and low-frequency species showed random associations. This co-occurrence pattern is similar to that reported by Lavender et al. (2019), who showed that most pairwise co-occurrence tests perform best for medium prevalence species. This specific abundance region is characterized by key indicator species, habitat specialists and generalists with clustered distributions. Such strong co-occurrence profiles provide considerable evidence to support the argument that interspecific interaction patterns are a promising proxy measure for characterizing the distributions of medium prevalence species. Given that high-frequency species should be often co-occurring in the same plots, it is not clear why high-frequency species does not show stronger co-occurrence patterns. This leads to the conclusion that using species with stronger interaction profiles for composite-SDM avoid such uncertainties on inferred distributions. As stated by Tikhonov et al. (2017) most co-occurrence analysis relies on the assumption that the species associations are invariant in relation to the environment. Therefore, it is highly recommended to take account and minimize the influence of co-occurrence covariation in relation to the environment to reach precise SDM predictions based on co-occurrences.

Given the high degree of species co-occurrences documented here, I suspect that individual species level predictions should have limited use. Species are almost always found with other species in a community and standalone inferences ignore interspecies relations. If community

assembly mechanisms are a result of deterministic abiotic gradients and species interactions (Hutchinson 1978, Drake 1990), species segregation and co-location can be inferred from observed co-occurrence patterns (Ulrich 2004, Veech 2013, Thuiller et al. 2015, Ulrich et al. 2017). The composite-SDM modelling framework is conceptualized to assess the possibilities of such processes, and to use those to reduce uncertainties in SDMs. This approach assumes observed species presences and inferred strength of co-occurrence patterns are a result of ecological processes in response to abiotic environmental variates. The research does not indicate any consistent kappa coefficient increase or decrease across the range of low to high-frequency species between the baseline direct reflectance models and the composite-SDM framework. In theoretical terms, the kappa coefficient goes up when there is a minimum of misclassifications from both presences and absences. In the proposed methodology, I clearly noticed that decreases in false negatives (increased correctly classified presences), increases false positives (decreased correctly classified absences) and vice-versa. It is not clear how error structure variations associated with each other at this moment and such influences on the composite-SDM framework does not perform any better compared to the baseline. The classification accuracy of predictions is mainly based on the binary map produced using a threshold chosen. Usually, increases or decreases of the threshold can influence the accuracy of results and finding the optimal solution is always context-specific. The consistent nature of accuracy changes along species distribution frequencies suggests that an examination of different thresholds might be an alternative path to optimize prediction accuracies of the composite framework. I separately averaged predicted probability surfaces of positive and negative co-occurring species to produce co-occurring geographical distributions of negative and positive locations for the target species. For example, maybe valuable to replace the averaging process of negative and positive co-occurring geo-spatial probabilities with alternative algorithms such as choosing the best principle components (Borcard et al. 2011, Legendre and Legendre 2012). Such algorithmic alternatives are likely to optimize geospatial co-occurrence probabilities for the target species and minimize inaccuracies in the composite-SDM process.

Although it is possible to infer co-occurrences using statistical methods such as distance-based ordinations (Legendre and Legendre 2012), pair-wise co-occurrence assessments (Veech 2014), or null-model randomization tests (Gotelli 2000), any method does not account for the strength of co-occurrence between species in consideration. An alternative solution is to construct latent variables based on both negative and positive co-occurrences. A latent construct represents

directly unmeasurable processes that might act as predictors to explain the response (Latimer et al. 2009, Tikhonov et al. 2017, Tobler et al. 2019), in this case, the relationship between target species occurrence in relation to co-occurring interspecific relationships. Co-occurring species frequencies could be entered as a covariate or weighted variable influencing the development of the latent factors. Another issue in most of the statistical methods discussed above is that the methods heavily depend on the assumption that spatial and temporal species associations are not influenced by environmental gradients. Recent research shows there are challenges in interpreting co-occurrences mainly because the pattern might be the result of either ecological interactions or a response to abiotic variates. In general terms, environmental co-variates might be responsible for determining either the occurrence of a species or the co-occurrence with others or both. Analysing such dependencies of species associations with abiotic-environmental co-variates is a recent development to deal with challenges of this assumption (Tikhonov et al. 2017). The method proposed by Tikhonov et al. (2017) considers environmental variables that influence species associations as latent factors and species associations are a result of latent factor loadings that reflect the responses of the species to the latent variable. The latent variable technique proposed has the potential to produce better results by incorporating species interrelationships with their associations to the environment.

5.5 Technical limitations

I used the BIOMOD-2 R package (Wilfried Thuiller et al. 2016, R Core Team 2019) to train all models. The programme was selected on the basis of capabilities to handle various SDM algorithms within one platform. Furthermore, the library is capable of producing very useful accuracy measures such as ROC, Kappa and TSS are very useful among many others available. The training process was time-consuming due to processing challenges including handling high-resolution raster data in the R software environment. The most pragmatic solution was to convert raster information into tables with the cost of higher file size. The result was very large tables of data files in comparison to the same data in the raster format; processing such large tables demanded very high capacity computing capabilities. I used the advanced computing infrastructure: large memory server “Meton” which is a Linux server with 2 terabytes (= 2048 gigabytes) of random-access memory. BIOMOD-2 package was used for model training and the trained mathematical formulae were manually transferred to ArcGIS software; manual transfer

was necessary due to integration incompatibilities between the R and ArcGIS software packages. The process would be sped up substantially if there were an ability to call BIOMOD-2 model results within ArcGIS.

5.6 Application challenges, and new directions

Species distribution models (SDM) with remotely sensed (RS) imagery are a widely used tool for ecological studies and conservation planning but inherited with various challenges. Issues including small plant size, small numbers of observations, and scattered distribution patterns due to differences in occupying favourable environmental gradients (Santika 2011, Zurell et al. 2016) are more pronounced for grassland related applications and understory species in comparison to communities with larger target species such as forests. A wide array of potential solutions to such issues have been recently published to facilitate biogeographical processes mapping including SDMs (Zimmermann et al. 2007, Aragón and Oesterheld 2008, Wang and Qu 2009, Lechner et al. 2012, He et al. 2015, Rocchini et al. 2015). The primary contribution from my thesis is the method to link species ecological relationships (co-occurrences) with remotely sensed (RS) predictors to better predict species occupancy patterns. Use of RS data provides a great opportunity to explore spatio-temporal scales which are impossible to incorporate otherwise. The wider coverage areas, longer time-lapse, and low-cost availability of data sources are factors for the rapid expansion of applications using RS data sources.

It is possible to obtain relatively accurate details of species distribution patterns using direct predictors such as soil moisture, nutrient distribution pattern, litter content, and soil temperature variations (Austin et al. 1984, Austin and Meyers 1996, Miller and Franklin 2002, Guisan and Thuiller 2005, Elith et al. 2006, Elith and Leathwick 2009, Thuiller 2013). However, such methods are limited to smaller geographical areas due to difficulties obtaining maps of predictor variables in contrast to the use of image reflectance bands as predictors to explain ecologically functional relationships. Usually, remotely sensed images accurately capture biophysical factors that are easy to link to the ecological niche space of the target species (Elith and Graham 2009, Cord et al. 2013, He et al. 2015, Rocchini et al. 2015). Often, however, the spatial scale of ecologically meaningful predictors does not correspond to either RS data spatial scales or field-based data that researchers collect for their analysis. Minimizing such disparities are crucial to achieving higher predictive

performance of RS based SDM systems. Reflectance data are usually an indirect measure of biophysical properties. Sometimes, it is possible to experience spatio-temporal non-stationarity between species occurrences and reflectance-based predictors coming from different temporal scales and different sensor platforms. It is therefore essential to develop pragmatic methods to bridge the gap between plant species biophysical properties driven by environmental drivers and RS data. Measurement of leaf area index (LAI), fraction of absorbed photosynthetically active radiation (fPAR), and land surface temperature (LST) is very good proxies of plant and environmental biophysical properties. Although such parameters are difficult to measure, it provides greater details of energy levels of the environment and water distribution patterns contribute significantly to shape species existence in the landscape.

I discussed throughout my thesis how prediction confidence goes down both with low numbers of presences and very high numbers of presences. It should be clearly stated that low frequency species are not always going to produce less confident results with RS images. Usually, prediction confidence increases when the species occurs in distinct habitats where it is visually and spectrally distinct from the surroundings, even if the frequency of occurrence is low. Such distinct habitat features are often controlled by direct environmental gradients that favour a specific set of species or a larger-bodied indicator species that has full control over its microclimate. Both situations are likely to produce spectrally distinct features in remotely acquired images. It is apparent, however, that prediction confidence is low when a species is of low frequency or when the species widespread many different habitats. This presents a significant challenge for small-bodied grassland species compared to larger plant species such as trees that generate distinct spectral features in the image. Low-frequency grassland plants are generally unlikely to produce such distinct features in the image as the target species is a part of different plant combinations distributed across the landscape. Therefore, the target species occurrences are associated with highly variable reflectance patterns. When the species is low frequency, there are a proportionately high number of absences associated with reflectance where presences are observed. Such associations are likely to reduce optimal algorithmic performance by producing more false positives and false negatives that will finally minimize overall prediction confidence. The best method to deal with such species is to understand the distribution pattern and habitat affinities at the initial stage of modelling. If the species is found across various habitats, it is highly recommended to segment the image to analyze species distributions in relation to different

reflectance regions of the image. Such an analysis is likely to provide insights into the degree of uncertainty that should be expected from the prediction. The uncertainty is a result of observed absences in similar habitats where few presences were reported. The analyst should be able to justify higher false positives as those locations likely represent potential, but unoccupied, habitat for such species.

I found that very high-frequency species also suffer from a low predictive performance. These species, such as *Carex* at Kernen Prairie, are often small-bodied sub-canopy species present under a wide variety of dominant grass or shrub species. The smaller number of absences is the obvious cause for such uncertain results. However, I suspect, high predictor variability associated with the observed presences also another contributor to low predictive confidence. The higher predictor variability is inevitable when the species is wider spread and occurs with various other species organizations of the landscape and likely to reduces reflectance uniqueness associated with presences as similar reflectance levels most likely to associated with absences as well. This is a unique example indicates the importance of absences and associated predictors compared to presences to achieve higher predictive performance in species modelling. Difficulties with these species were not only apparent with GLM, but also with other algorithms such as GAM, RF, MARS and ANN. There is no literature examining how to predict species with nearly 100 percent observed presence in the sample space. One could argue that this is a trivial case with nothing to predict if the species is occurring everywhere. However, cases of 90 percent or more presences are possible, and are likely to produce highly uncertain predictions. The only practical solution for this problem is the expansion of the sampling strategy to bring balance between data likely contribute positively to improve prediction accuracies.

I would like to conclude my discussion restating one of the key take-home messages from the study. The study was formulated assuming that the algorithmic sophistication has a positive influence on the predictive performance, that object-based image clustering methodologies have an ability to bring out plant community structural organizations to positively contribute to model accuracy, and co-occurring species in the composite-SDM framework will have a significant positive influence on predictive performance. The biggest finding surprised me was that the GLMs outperformed other complex algorithms with no sacrifice of accuracy for simplicity. Optimal performance at medium levels of observed species frequencies and marginal performance at both

low and high-frequency distributions suggest that the algorithmic performance is maximum where presence-absence balance is reached and distributions are associated with unique features of predictors available to attribute existing prevalence. Therefore, it is highly recommended to invest the necessary resources at the planning stage to identify appropriate variables to attribute species existence compared to non-existence and develop sampling strategy to reach balanced observations in comparison to the total sample frame. In my opinion, it is not species rarity itself that causes problems, rather such predictions are uncertain due to presences and absences both being attributed to similar predictor characteristics. This research domain is a new avenue to explore as there is no literature available clearly explaining how presence-absence balance impacts on prediction certainty. It is important to evaluate the influence in relation to predictor feature variation. The composite-SDM framework takes into account the total plant community that the species co-exist, however, the framework was tested on the assumption that the spatial and temporal observed co-occurrence is constant. The influence of abiotic variates on the occurrence and co-occurrence of species for given space and time is most likely to invalidate constant co-occurrence assumption. Inclusion of the reality of species co-existence in relation to interspecific relationships with other species and abiotic environment is interesting, the development of such interaction modelling framework is very time and labour intensive. Therefore, the development of automation techniques to easily handle such a high number of species and variables are critical to consider in future work.

References

Al-Masrahy, M. A., and N. P. Mountney. 2013. Remote sensing of spatial variability in aeolian dune and interdune morphology in the Rub' Al-Khali, Saudi Arabia. Aeolian Research **11**:155-170.

Allouche, O., A. Tsoar, and R. Kadmon. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology **43**:1223-1232.

Anderson, R. P. 2017. When and how should biotic interactions be considered in models of species niches and distributions? Journal of Biogeography **44**:8-17.

Aragón, R., and M. Oesterheld. 2008. Linking vegetation heterogeneity and functional attributes of temperate grasslands through remote sensing. Applied Vegetation Science **11**:117-130.

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

Argus, G. W., and J. W. Steele. 1979. A Reevaluation of the Taxonomy of Salix tyrrellii, a Sand Dune Endemic. Systematic Botany **4**:163-177.

Ashcroft, M. B., D. H. King, B. Raymond, J. D. Turnbull, J. Wasley, and S. A. Robinson. 2017. Moving beyond presence and absence when examining changes in species distributions. Global Change Biology:n/a-n/a.

Attanayake, A. U., D. Xu, X. Guo, and E. G. Lamb. 2018. Long-term sand dune spatio-temporal dynamics and endemic plant habitat extent in the Athabasca sand dunes of northern Saskatchewan. Remote Sensing in Ecology and Conservation **5**:70-86.

Austin, M. P., L. Belbin, J. A. Meyers, M. D. Doherty, and M. Luoto. 2006. Evaluation of statistical models used for predicting plant species distributions: Role of artificial data and theory. Ecological Modelling **199**:197-216.

Austin, M. P., R. B. Cunningham, and P. M. Fleming. 1984. New approaches to direct gradient analysis using environmental scalars and statistical curve-fitting procedures. Vegetatio **55**:11-27.

Austin, M. P., and M. J. Gaywood. 1994. Current problems of environmental gradients and species response curves in relation to continuum theory. Journal of Vegetation Science **5**:473-482.

Austin, M. P., and J. A. Meyers. 1996. Current approaches to modelling the environmental niche of eucalypts: implication for management of forest biodiversity. Forest Ecology and Management **85**:95-106.

Austin, M. P., A. O. Nicholls, and C. R. Margules. 1990. Measurement of the Realized Qualitative Niche: Environmental Niches of Five Eucalyptus Species. Ecological Monographs **60**:161-177.

Barbet-Massin, M., F. Jiguet, C. H. Albert, and W. Thuiller. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? Methods in Ecology and Evolution **3**:327-338.

Bean, W. T., R. Stafford, and J. S. Brashares. 2012. The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. Ecography **35**:250-258.

Blaschke, T. 2010. Object based image analysis for remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing **65**:2-16.

Blaschke, T., G. J. Hay, M. Kelly, S. Lang, P. Hofmann, E. Addink, R. Queiroz Feitosa, F. van der Meer, H. van der Werff, F. van Coillie, and D. Tiede. 2014. Geographic Object-Based Image Analysis – Towards a new paradigm. ISPRS Journal of Photogrammetry and Remote Sensing **87**:180-191.

Blaschke, T., S. Lang, and G. Hay. 2008. Object-Based Image Analysis: Spatial Concepts for Knowledge-Driven Remote Sensing Applications. Springer Berlin Heidelberg.

Borcard, D., F. Gillet, and P. Legendre. 2011. Numerical Ecology with R. Springer New York.

Bradley, B. A., A. D. Olsson, O. Wang, B. G. Dickson, L. Pelech, S. E. Sesnie, and L. J. Zachmann. 2012. Species detection vs. habitat suitability: Are we biasing habitat suitability models with remotely sensed data? Ecological Modelling **244**:57-64.

Breiman, L. 1996. Bagging Predictors. Machine Learning **24**:123-140.

Breiman, L. 2001. Using Iterated Bagging to Debias Regressions. Machine Learning **45**:261-277.

Brooker, R. W., F. T. Maestre, R. M. Callaway, C. L. Lortie, L. A. Cavieres, G. Kunstler, P. Liancourt, K. Tielbörger, J. M. J. Travis, F. Anthelme, C. Armas, L. Coll, E. Corcket, S. Delzon, E. Forey, Z. Kikvidze, J. Olofsson, F. Pugnaire, C. L. Quiroz, P. Saccone, K. Schiffers, M. Seifan, B. Touzard, and R. Michalet. 2008. Facilitation in plant communities: the past, the present, and the future. Journal of Ecology **96**:18-34.

Buckley, H. L., B. S. Case, and A. M. Ellison. 2016. Using codispersion analysis to characterize spatial patterns in species co-occurrences. Ecology **97**:32-39.

Campbell, J. B., and R. H. Wynne. 2011. Introduction to Remote Sensing, Fifth Edition. Guilford Publications.

Carson, M. A., and P. A. MacLean. 1986. Development of hybrid aeolian dunes: the William River dune field, northwest Saskatchewan, Canada. Canadian Journal of Earth Sciences **23**:1974-1990.

Cassini, M. H. 2011. Ecological principles of species distribution models: the habitat matching rule. Journal of Biogeography **38**:2057-2065.

Cazelles, K., M. Araújo, N. Mouquet, and D. Gravel. 2015. A theory for species co-occurrence in interaction networks. Theoretical Ecology:1-10.

Chander, G., M. O. Haque, E. Micijevic, and J. A. Barsi. 2008. Landsat 5 Thematic Mapper (TM) Recalibration Procedure for Data Processed using the National Landsat Archive Production System (NLAPS). Pages IV - 1360-IV - 1363 *in* Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International.

Chander, G., B. L. Markham, and D. L. Helder. 2009. Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors. Remote Sensing of Environment **113**:893-903.

Chase, J. M., and M. A. Leibold. 2003. Ecological Niches: Linking Classical and Contemporary Approaches. University of Chicago Press.

Coburn, C. A., and A. C. B. Roberts. 2004. A multiscale texture analysis procedure for improved forest stand classification. International Journal of Remote Sensing **25**:4287-4308.

Cole, B., J. McMorrow, and M. Evans. 2014. Empirical Modelling of Vegetation Abundance from Airborne Hyperspectral Data for Upland Peatland Restoration Monitoring. Remote Sensing **6**:716-739.

Colwell, R. K., and T. F. Rangel. 2009. Hutchinson's duality: The once and future niche. Proceedings of the National Academy of Sciences **106**:19651-19658.

Cooper, R. L., and D. D. Cass. 2003. A comparative epidermis study of the Athabasca sand dune willows (Salix; Salicaceae) and their putative progenitors. Canadian Journal of Botany **81**:749-754.

Cord, A., and D. Rödder. 2011. Inclusion of habitat availability in species distribution models through multi-temporal remote sensing data? Ecological Applications **21**:3285-3298.

Cord, A. F., D. Klein, D. S. Gernandt, J. A. P. de la Rosa, and S. Dech. 2014a. Remote sensing data can improve predictions of species richness by stacked species distribution models: a case study for Mexican pines. Journal of Biogeography **41**:736-748.

Cord, A. F., D. Klein, F. Mora, and S. Dech. 2014b. Comparing the suitability of classified land cover data and remote sensing variables for modeling distribution patterns of plants. Ecological Modelling **272**:129-140.

Cord, A. F., R. K. Meentemeyer, P. J. Leitão, and T. Václavík. 2013. Modelling species distributions with remote sensing data: bridging disciplinary perspectives. Journal of Biogeography **40**:2226-2227.

COSEWIC. 2018. COSEWIC Assessment and Status Report on the Athabasca Endemics Large-headed Woolly Yarrow (Achillea millefolium var. megacephala) Athabasca Thrift (Armeria maritima ssp. interior) Mackenzie Hairgrass (Deschampsia mackenzieana) Sand-dune Short-capsuled Willow (Salix brachycarpa var. psammophila) Turnor's Willow (Salix turnorii) Blanket-leaved Willow (Salix silicicola) Floccose Tansy (Tanacetum huronense var. floccosum) in Canada 2018. Committee on the status of Endangered Wildlife in Canada (COSEWIC) Special concern 2018.

Coupland, R. T., and T. C. Brayshaw. 1953. The Fescue Grassland in Saskatchewan. Ecology **34**:386-405.

Crisci, C., B. Ghattas, and G. Perera. 2012. A review of supervised machine learning algorithms and their applications to ecological data. Ecological Modelling **240**:113-122.

D'Amen, M., C. Rahbek, N. E. Zimmermann, and A. Guisan. 2015. Spatial predictions at the community level: from current approaches to future frameworks. Biological Reviews:n/a-n/a.

Dashtekian, E. S. K. 2013. Analysis of land use-land covers changes using normalized difference vegetation index (NDVI) differencing and classification methods. African Journal of Agricultural Research **Vol.8**:4614-4622.

Drake, J. A. 1990. The mechanics of community assembly and succession. Journal of Theoretical Biology **147**:213-233.

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

Dubuis, A., J. Pottier, V. Rion, L. Pellissier, J.-P. Theurillat, and A. Guisan. 2011. Predicting spatial patterns of plant species richness: a comparison of direct macroecological and species stacking modelling approaches. Diversity and Distributions **17**:1122-1131.

Elith, J., and C. H. Graham. 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. Ecography **32**:66-77.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. McC. M. Overton, A. Townsend Peterson, S. J. Phillips, K. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberón, S. Williams, M. S. Wisz, and N. E. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography **29**:129-151.

Elith, J., and J. R. Leathwick. 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics **40**:677-697.

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

Elith, J., S. J. Phillips, T. Hastie, M. Dudík, Y. E. Chee, and C. J. Yates. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions **17**:43-57.

Engler, R., A. Guisan, and L. Rechsteiner. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal of Applied Ecology **41**:263-274.

Ewing, R. C., and G. Kocurek. 2010. Aeolian dune-field pattern boundary conditions. Geomorphology **114**:175 - 187.

Ewing, R. C., G. Kocurek, and L. W. Lake. 2006. Pattern analysis of dune-field parameters. Earth Surface Processes and Landforms **31**:1176-1191.

Ferrier, S., and A. Guisan. 2006. Spatial modelling of biodiversity at the community level. Journal of Applied Ecology **43**:393-404.

Fielding, A. H., and j. F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation **24**:38-49.

Flora of North America Editorial Committee, e. 1993+. Flora of North America North of Mexico. New York and Oxford. **20+ vols**.

Franklin, J., and J. A. Miller. 2009. Mapping Species Distributions: Spatial Inference and Prediction. Cambridge University Press.

Franklin, J., K. E. Wejnert, S. A. Hathaway, C. J. Rochester, and R. N. Fisher. 2009. Effect of species rarity on the accuracy of species distribution models for reptiles and amphibians in southern California. Diversity and Distributions **15**:167-177.

Freeman, E. A., and G. G. Moisen. 2008. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. Ecological Modelling **217**:48-58.

Friedman, J. H., and C. B. Roosen. 1995. An introduction to multivariate adaptive regression splines. Statistical Methods in Medical Research **4**:197-217.

Gaston, K. J. 1996. Species-range-size distributions: patterns, mechanisms and implications. Trends in Ecology & Evolution **11**:197-201.

Gaston, K. J., T. M. Blackburn, J. J. D. Greenwood, R. D. Gregory, R. M. Quinn, and J. H. Lawton. 2000. Abundance–occupancy relationships. Journal of Applied Ecology **37**:39-59.

Gaston, K. J., and J. H. Lawton. 1990. Effects of Scale and Habitat on the Relationship between Regional Distribution and Local Abundance. Oikos **58**:329-335.

Gauch, H. G., Jr., and R. H. Whittaker. 1981. Hierarchical Classification of Community Data. Journal of Ecology **69**:537-557.

Gogol-Prokurat, M. 2011. Predicting habitat suitability for rare plants at local spatial scales using a species distribution model. Ecological Applications **21**:33-47.

Gotelli, N. J. 2000. NULL MODEL ANALYSIS OF SPECIES CO-OCCURRENCE PATTERNS. Ecology **81**:2606-2621.

Government of Canada. 2015. Canadian Weather. Government of Canada.

Griffith, D. M., J. A. Veech, and C. J. Marsh. 2016. cooccur: Probabilistic Species Co-Occurrence Analysis in R. Journal of Statistical Software **69**:17.

Grinnell, J. 1917. The Niche-Relationships of the California Thrasher. The Auk **34**:427-433.

Guillera-Arroita, G. 2017. Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. Ecography **40**:n/a-n/a.

Guillera-Arroita, G., J. J. Lahoz-Monfort, J. Elith, A. Gordon, H. Kujala, P. E. Lentini, M. A. McCarthy, R. Tingley, and B. A. Wintle. 2015. Is my species distribution model fit for purpose? Matching data and models to applications. Global Ecology and Biogeography **24**:276-292.

Guisan, A., O. Broennimann, R. Engler, M. Vust, N. G. Yoccoz, A. Lehmann, and N. E. Zimmermann. 2006a. Using Niche-Based Models to Improve the Sampling of Rare Species. Conservation Biology **20**:501-511.

Guisan, A., T. C. Edwards Jr, and T. Hastie. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling **157**:89-100.

Guisan, A., A. Lehmann, S. Ferrier, M. Austin, J. M. C. Overton, R. Aspinall, and T. Hastie. 2006b. Making better biogeographical predictions of species' distributions. Journal of Applied Ecology **43**:386-392.

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

Guisan, A., and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecological Modelling **135**:147-186.

Guy, A. L., J. M. Mischkolz, and E. G. Lamb. 2013. Limited effects of simulated acidic deposition on seedling survivorship and root morphology of endemic plant taxa of the Athabasca Sand Dunes in well-watered greenhouse trials. Botany **91**:176-181.

Hanski, I. 1998. Metapopulation dynamics. Nature **396**:41-49.

Haralick, R. M., K. Shanmugam, and I. Dinstein. 1973. Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics **SMC-3**:610-621.

Hastie, T., and R. Tibshirani. 1990. Generalized Additive Models. Chapman and Hall.

He, K. S., B. A. Bradley, A. F. Cord, D. Rocchini, M.-N. Tuanmu, S. Schmidtlein, W. Turner, M. Wegmann, and N. Pettorelli. 2015. Will remote sensing shape the next generation of species distribution models? Remote Sensing in Ecology and Conservation **1**:4-18.

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

Hesse, R. 2009. Using remote sensing to quantify aeolian transport and estimate the age of the terminal dune field Dunas Pampa Blanca in southern Peru. Quaternary Research **71**:426-436.

Hirzel, A., and A. Guisan. 2002. Which is the optimal sampling strategy for habitat suitability modelling. Ecological Modelling **157**:331-341.

Hirzel, A. H., and R. Arlettaz. 2003. Modeling Habitat Suitability for Complex Species Distributions by Environmental-Distance Geometric Mean. Environmental Management **32**:614-623.

Hirzel, A. H., G. Le Lay, V. Helfer, C. Randin, and A. Guisan. 2006. Evaluating the ability of habitat suitability models to predict species presences. Ecological Modelling **199**:142-152.

Hjort, J., and M. Marmion. 2008. Effects of sample size on the accuracy of geomorphological models. Geomorphology **102**:341-350.

Holben, B. N. 1986. Characteristics of maximum-value composite images from temporal AVHRR data. International Journal of Remote Sensing **7**:1417-1434.

Holt, R. D. 2009. Bringing the Hutchinsonian Niche into the 21st Century: Ecological and Evolutionary Perspectives. Proceedings of the National Academy of Sciences of the United States of America **106**:19659-19665.

Howari, F. M., A. Baghdady, and P. C. Goodell. 2007. Mineralogical and gemorphological characterization of sand dunes in the eastern part of United Arab Emirates using orbital remote sensing integrated with field investigations. Geomorphology **83**:67 - 81.

Hugenholtz, C. H. 2005a. Biogeomorphic model of dunefield activation and stabilization on the northern Great Plains. Geomorphology **70**:53 - 70.

Hugenholtz, C. H. 2005b. Recent stabilization of active sand dunes on the Canadian prairies and relation to recent climate variations. Geomorphology **68**:131 - 147.

Hugenholtz, C. H., N. Levin, T. E. Barchyn, and M. C. Baddock. 2012. Remote sensing and spatial analysis of aeolian sand dunes: A review and outlook. Earth-Science Reviews **111**:319-334.

Hutchinson, G. E. 1957. Concluding remarks. Population Studies: Animal Ecology and Demography. Cold Spring Harbor Symposium on Quantitative Biology **22**:415-457.

Hutchinson, G. E. 1959. Homage to Santa Rosalia or Why Are There So Many Kinds of Animals? The American Naturalist **93**:145-159.

Hutchinson, G. E. 1978. An Introduction to Population Ecology. Yale University Press.

Izco, J. 1998. Types of rarity of plant communities. Journal of Vegetation Science **9**:641-646.

Jensen, J. R. 2005. Introductory digital image processing: a remote sensing perspective. Pearson Prentice Hall.

Jiménez-Valverde, A., J. Lobo, and J. Hortal. 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. Community Ecology **10**:196-205.

Jiménez-Valverde, A., and J. M. Lobo. 2007. Threshold criteria for conversion of probability of species presence to either–or presence–absence. Acta Oecologica **31**:361-369.

Kruckeberg, A. R., and D. Rabinowitz. 1985. Biological Aspects of Endemism in Higher Plants. Annual Review of Ecology and Systematics **16**:447-479.

Kunin, W. E., and K. Gaston. 1996. The Biology of Rarity: Causes and consequences of rare—common differences. Springer Netherlands.

Laliberte, A. S., and A. Rango. 2009. Texture and Scale in Object-Based Analysis of Subdecimeter Resolution Unmanned Aerial Vehicle (UAV) Imagery. IEEE Transactions on Geoscience and Remote Sensing **47**:761-770.

Lamb, E. G., and D. D. Guedo. 2012. The distribution, abundance, and environmental affinities of the endemic vascular plant taxa of the athabasca sand dunes of northern saskatchewan. Ecoscience **19**:161-169.

Lamb, E. G., J. M. Mischkolz, and D. Guedo. 2011. The distribution and abundance of the endemic vascular plant taxa of the Athabasca Sand Dunes of northern Saskatchewan. University of Saskatchewan, Department of Plant Science.

Lamb, E. G., T. Winsley, C. L. Piper, S. A. Freidrich, and S. D. Siciliano. 2016. A high-throughput belowground plant diversity assay using next-generation sequencing of the trnL intron. Plant and Soil **404**:361-372.

Lancaster, N. 1995. Geomorphology of Desert Dunes. Routledge.

Landis, J. R., and G. G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. Biometrics **33**:159-174.

Latimer, A. M., S. Banerjee, H. Sang Jr, E. S. Mosher, and J. A. Silander Jr. 2009. Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States. Ecology Letters **12**:144-154.

Lavender, T. M., B. S. Schamp, S. E. Arnott, and J. A. Rusak. 2019. A comparative evaluation of five common pairwise tests of species association. Ecology **100**:e02640.

Leathwick, J. R., J. Elith, and T. Hastie. 2006. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. Ecological Modelling **199**:188-196.

Lechner, A., W. Langford, S. Bekessy, and S. Jones. 2012. Are landscape ecologists addressing uncertainty in their remote sensing data? Landscape Ecology **27**:1249-1261.

Legendre, P., and L. Legendre. 2012. Numerical Ecology. Elsevier Science Limited.

Levin, N., E. Ben-Dor, and A. Karnieli. 2004. Topographic information of sand dunes as extracted from shading effects using Landsat images. Remote Sensing of Environment **90**:190-209.

Liu, C., P. M. Berry, T. P. Dawson, and R. G. Pearson. 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography **28**:385-393.

Liu, C., M. White, and G. Newell. 2013. Selecting thresholds for the prediction of species occurrence with presence-only data. Journal of Biogeography **40**:778-789.

Livingstone, I., G. F. S. Wiggs, and C. M. Weaver. 2007. Geomorphology of desert sand dunes: A review of recent progress. Earth-Science Reviews **80**:239 - 257.

Lobo, J. M., and M. F. Tognelli. 2011. Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. Journal for Nature Conservation **19**:1-7.

Macdonald, S. E., C. C. Chinnappa, D. M. Reid, and B. G. Purdy. 1987. Population differentiation of the Stellaria longipes complex within Saskatchewan's Athabasca sand dunes. Canadian Journal of Botany **65**:1726-1732.

Manel, S., H. C. Williams, and S. J. Ormerod. 2001. Evaluating presence–absence models in ecology: the need to account for prevalence. Journal of Applied Ecology **38**:921-931.

Markogianni, V., E. Dimitriou, and D. P. Kalivas. 2012. Land-use and vegetation change detection in Plastira artificial lake catchment (Greece) by using remote-sensing and GIS techniques. International Journal of Remote Sensing **34**:1265-1281.

Mateo, R. G., Á. M. Felicísimo, and J. Muñoz. 2010. Effects of the number of presences on reliability and stability of MARS species distribution models: the importance of regional niche variation and ecological heterogeneity. Journal of Vegetation Science **21**:908-922.

McKee, E. D. 1979. A study of global sand seas. USGS, USGS Numbered Series - Professional Paper.

McNickle, G. G., E. G. Lamb, M. Lavender, J. F. Cahill, B. S. Schamp, S. D. Siciliano, R. Condit, S. P. Hubbell, and J. L. Baltzer. 2018. Checkerboard score–area relationships reveal spatial scales of plant community structure. Oikos **127**:415-426.

McPherson, J. M., W. Jetz, and D. J. Rogers. 2006. Using coarse-grained occurrence data to predict species distributions at finer spatial resolutions—possibilities and limitations. Ecological Modelling **192**:499-522.

Merow, C., M. J. Smith, and J. A. Silander. 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography:no-no.

Miller, J. 2010. Species Distribution Modeling. Geography Compass **4**:490-509.

Miller, J., and J. Franklin. 2002. Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. Ecological Modelling **157**:227-247.

Mohamed, I. N. L., and G. Verstraeten. 2012. Analyzing dune dynamics at the dune-field scale based on multi-temporal analysis of Landsat-TM images. Remote Sensing of Environment **119**:105-117.

Neeson, T. M., and Y. Mandelik. 2014. Pairwise measures of species co-occurrence for choosing indicator species and quantifying overlap. Ecological Indicators **45**:721-727.

Neigh, C. S. R., C. J. Tucker, and J. R. G. Townshend. 2008. North American vegetation dynamics observed with multi-resolution satellite data. Remote Sensing of Environment **112**:1749 - 1772.

Norberg, A., N. Abrego, F. G. Blanchet, F. R. Adler, B. J. Anderson, J. Anttila, M. B. Araújo, T. Dallas, D. Dunson, J. Elith, S. D. Foster, R. Fox, J. Franklin, W. Godsoe, A. Guisan, B. O'Hara, N. A. Hill, R. D. Holt, F. K. C. Hui, M. Husby, J. A. Kålås, A. Lehikoinen, M. Luoto, H. K. Mod, G. Newell, I. Renner, T. Roslin, J. Soininen, W. Thuiller, J. Vanhatalo, D. Warton, M. White, N. E. Zimmermann, D. Gravel, and O. Ovaskainen. 2019. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. Ecological Monographs **0**:e01370.

Norwine, J., and D. H. Greegor. 1983. Vegetation classification based on advanced very high resolution radiometer (AVHRR) satellite imagery. Remote Sensing of Environment **13**:69 - 87.

Okin, G. S., and T. H. Painter. 2004. Effect of grain size on remotely sensed spectral reflectance of sandy desert surfaces. Remote Sensing of Environment **89**:272-280.

Olden, Julian D., Joshua J. Lawler, and N. L. Poff. 2008. Machine Learning Methods Without Tears: A Primer for Ecologists. The Quarterly Review of Biology **83**:171-193.

Özyavuz, M. 2010. Analysis of Changes in Vegetation Using Multitemporal Satellite Imagery, the Case of Tekirdağ Coastal Town. Journal of Coastal Research **26**:1038-1046.

Paisley, E. C. I., N. Lancaster, L. R. Gaddis, and R. Greeley. 1991. Discrimination of active and inactive sand from remote sensing: Kelso dunes, Mojave desert, California. Remote Sensing of Environment **37**:153-166.

Parker Gay Jr, S. 1999. Observations regarding the movement of barchan sand dunes in the Nazca to Tanaca area of southern Peru. Geomorphology **27**:279-293.

Pearson, R. G., C. J. Raxworthy, M. Nakamura, and A. Townsend Peterson. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeography **34**:102-117.

Phillips, S. 2012. Inferring prevalence from presence-only data: a response to 'Can we model the probability of presence of species without absence data?'. Ecography **35**:385-387.

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

Phillips, S. J., M. Dudík, J. Elith, C. H. Graham, A. Lehmann, J. Leathwick, and S. Ferrier. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications **19**:181-197.

Pineda, E., and J. M. Lobo. 2009. Assessing the accuracy of species distribution models to predict amphibian species richness patterns. Journal of Animal Ecology **78**:182-190.

Pollock, L. J., R. Tingley, W. K. Morris, N. Golding, R. B. O'Hara, K. M. Parris, P. A. Vesk, and M. A. McCarthy. 2014. Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). Methods in Ecology and Evolution **5**:397-406.

Pulliam, H. R. 1988. Sources, Sinks, and Population Regulation. The American Naturalist **132**:652-661.

Pulliam, H. R. 2000. On the relationship between niche and distribution. Ecology Letters **3**:349-361.

Pulliam, H. R., and B. J. Danielson. 1991. Sources, Sinks, and Habitat Selection: A Landscape Perspective on Population Dynamics. The American Naturalist **137**:S50-S66.

Purdy, B. G., and R. J. Bayer. 1995. Genetic Diversity in the Tetraploid Sand Dune Endemic Deschampsia mackenzieana and its Widespread Diploid Progenitor D. cespitosa (Poaceae). American Journal of Botany **82**:121-130.

Purdy, B. G., and R. J. Bayer. 1996. Genetic variation in populations of the endemic Achillea millefolium ssp. megacephala from the Athabasca sand dunes and the widespread ssp. lanulosa in western North America. Canadian Journal of Botany **74**:1138-1146.

Pylypec, B. 1986. The Kernen Prairie – a relict fescue grassland near Saskatoon, Saskatchewan. Blue Jay **44**:222-231.

R Core Team. 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team, R. F. f. S. C. 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Raup, H. M. 1936. Phytogeographical studies in the Athabasca-Great Slave Lake region. Journal of the Arnold Arboretum **17**.

Raup, H. M., and G. W. Argus. 1982. The Lake Athabasca sand dunes of northern Saskatchewan and Alberta, Canada. National Museums of Canada, Ottawa.

Richards, J. A., and X. Jia. 1999. Remote Sensing Digital Image Analysis: An Introduction. Springer.

Rocchini, D., D. S. Boyd, J.-B. Féret, G. M. Foody, K. S. He, A. Lausch, H. Nagendra, M. Wegmann, and N. Pettorelli. 2015. Satellite remote sensing to monitor species diversity: potential and pitfalls. Remote Sensing in Ecology and Conservation:n/a-n/a.

Rodríguez-Rey, M., A. Jiménez-Valverde, and P. Acevedo. 2013. Species distribution models predict range expansion better than chance but not better than a simple dispersal model. Ecological Modelling **256**:1-5.

Romo, J. T. 2003. Reintroducing fire for conservation of fescue prairie association remnants in the northern Great Plains. Canadian Field Naturalist:89-99.

Roughgarden, J., S. W. Running, and P. A. Matson. 1991. What Does Remote Sensing Do For Ecology? Ecology **72**:1918-1922.

Ryherd, S., and C. E. Woodcock. 1996. Combining spectral and texture data in the segmentation of remotely sensed images. Photogrammetric engineering and remote sensing **62**:181-194.

Santika, T. 2011. Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. Global Ecology and Biogeography **20**:181-192.

Sheskin, D. J. 2003. Handbook of Parametric and Nonparametric Statistical Procedures: Third Edition. Taylor & Francis.

Slopek, J. I., and E. G. Lamb. 2017. Long-Term Efficacy of Glyphosate for Smooth Brome Control in Native Prairie. Invasive Plant Science and Management **10**:350-355.

Smith, M. O., S. L. Ustin, J. B. Adams, and A. R. Gillespie. 1990. Vegetation in deserts: I. A regional measure of abundance from multispectral images. Remote Sensing of Environment **31**:1-26.

Soberón, J. M. 2010. Niche and area of distribution modeling: a population ecology perspective. Ecography **33**:159-167.

Soulé, M. E. 1986. Conservation biology: the science of scarcity and diversity. Sinauer Associates.

Stockwell, D. R. B., and A. T. Peterson. 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling **148**:1-13.

Stokland, J. N., R. Halvorsen, and B. Støa. 2011. Species distribution modelling—Effect of design and sample size of pseudo-absence observations. Ecological Modelling **222**:1800-1809.

Strona, G., and J. A. Veech. 2015. A new measure of ecological network structure based on node overlap and segregation. Methods in Ecology and Evolution **6**:907-915.

Thorson, J. T., M. D. Scheuerell, A. O. Shelton, K. E. See, H. J. Skaug, and K. Kristensen. 2015. Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. Methods in Ecology and Evolution **6**:627-637.

Thuiller, W. 2003. BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change. Global Change Biology **9**:1353-1362.

Thuiller, W. 2013. On the importance of edaphic variables to predict plant species distributions - limits and prospects. Journal of Vegetation Science **24**:591-592.

Thuiller, W., L. Brotons, M. B. Araújo, and S. Lavorel. 2004. Effects of restricting environmental range of data to project current and future species distributions. Ecography **27**:165-172.

Thuiller, W., M. Guéguen, D. Georges, R. Bonet, L. Chalmandrier, L. Garraud, J. Renaud, C. Roquet, J. Van Es, N. E. Zimmermann, and S. Lavergne. 2014. Are different facets of plant diversity well protected against climate and land cover changes? A test study in the French Alps. Ecography **37**:1254-1266.

Thuiller, W., B. Lafourcade, R. Engler, and M. B. Araújo. 2009. BIOMOD – a platform for ensemble forecasting of species distributions. Ecography **32**:369-373.

Thuiller, W., S. Lavorel, M. T. Sykes, and M. B. Araújo. 2006. Using niche-based modelling to assess the impact of climate change on tree functional diversity in Europe. Diversity and Distributions **12**:49-60.

Thuiller, W., L. J. Pollock, M. Gueguen, and T. Münkemüller. 2015. From species distributions to meta-communities. Ecology Letters **18**:1321-1328.

Tikhonov, G., N. Abrego, D. Dunson, and O. Ovaskainen. 2017. Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. Methods in Ecology and Evolution **8**:443-452.

Tobler, M. W., M. Kéry, F. K. C. Hui, G. Guillera-Arroita, P. Knaus, and T. Sattler. 2019. Joint species distribution models with species correlations and imperfect detection. Ecology **100**:e02754.

Toutin, T. 2004. Review article: Geometric processing of remote sensing images: models, algorithms and methods. International Journal of Remote Sensing **25**:1893-1924.

Treitz, P. M., and P. J. Howarth. 1999. Hyperspectral remote sensing for estimating biophysical parameters of forest ecosystems. Progress in Physical Geography **23**:359-390.

Tsoar, H. 2004. Elongation and migration of sand dunes. Geomorphology **57**:293 - 302.

Tsoar, H. 2005. Sand dunes mobility and stability in relation to climate. Physica A: Statistical Mechanics and its Applications **357**:50 - 56.

U.S. Geological Survey. 2014. Maps, Imagery, and Publications. USGS, USGS.

U.S. Geological Survey. 2018. Maps, Imagery, and Publications. USGS, USGS.

Ulrich, W. 2004. Species co-occurrences and neutral models: reassessing J. M. Diamond's assembly rules. Oikos **107**:603-609.

Ulrich, W., and N. J. Gotelli. 2010. Null model analysis of species associations using abundance data. Ecology **91**:3384-3397.

Ulrich, W., and N. J. Gotelli. 2007. Disentangling community patterns of nestedness and species co-occurrence. Oikos **116**:2053-2061.

Ulrich, W., F. Jabot, and N. J. Gotelli. 2017. Competitive interactions change the pattern of species co-occurrences under neutral dispersal. Oikos **126**:91-100.

Veech, J. A. 2013. A probabilistic model for analysing species co-occurrence. Global Ecology and Biogeography **22**:252-260.

Veech, J. A. 2014. The pairwise approach to analysing species co-occurrence. Journal of Biogeography **41**:1029-1035.

Wang, L., and J. Qu. 2009. Satellite remote sensing applications for surface soil moisture monitoring: A review. Frontiers of Earth Science in China **3**:237-247.

Weng, Q. 2011. Advances in Environmental Remote Sensing: Sensors, Algorithms, and Applications. CRC Press.

Whittaker, R. H. 1965. Dominance and Diversity in Land Plant Communities. Science **147**:250-260.

Whittaker, R. H. 1967. Gradient analysis of vegetation*. Biological Reviews **42**:207-264.

Whittaker, R. H., S. A. Levin, and R. B. Root. 1973. Niche, Habitat, and Ecotope. The American Naturalist **107**:321-338.

Whittaker, R. H., S. A. Levin, and R. B. Root. 1975. On the Reasons for Distinguishing "Niche, Habitat, and Ecotope". The American Naturalist **109**:479-482.

Wilfried Thuiller, Damien Georges, Robin Engler, and F. Breiner. 2016. Ensemble Platform for Species Distribution Modeling. Page biomod2.

Williams, J. N., C. Seo, J. Thorne, J. K. Nelson, S. Erwin, J. M. O'Brien, and M. W. Schwartz. 2009. Using species distribution models to predict new occurrences for rare plants. Diversity and Distributions **15**:565-576.

Wisz, M., and A. Guisan. 2009. Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. BMC Ecology **9**:8.

Wisz, M. S., R. J. Hijmans, J. Li, A. T. Peterson, C. H. Graham, A. Guisan, and N. P. S. D. W. Group. 2008. Effects of sample size on the performance of species distribution models. Diversity and Distributions **14**:763-773.

Wolfe, S. A., D. J. Huntley, P. P. David, J. Ollerhead, D. J. Sauchyn, and G. M. MacDonald. 2001. Late 18th century drought-induced sand dune activity, Great Sand Hills, Saskatchewan. Canadian Journal of Earth Sciences **38**:105-117.

Wood, E. M., A. M. Pidgeon, V. C. Radeloff, and N. S. Keuler. 2012. Image texture as a remotely sensed measure of vegetation structure. Remote Sensing of Environment **121**:516-526.

Wood, S. N. 2006. Generalized Additive Models: An Introduction with R. Chapman & Hall/CRC.

Yao, Z. Y., T. Wang, Z. W. Han, W. M. Zhang, and A. G. Zhao. 2007. Migration of sand dunes on the northern Alxa Plateau, Inner Mongolia, China. Journal of Arid Environments **70**:80-93.

Yee, T. W., and M. Mackenzie. 2002. Vector generalized additive models in plant ecology. Ecological Modelling **157**:141-156.

Yee, T. W., and N. D. Mitchell. 1991. Generalized Additive Models in Plant Ecology. Journal of Vegetation Science **2**:587-602.

Young, N. E., R. S. Anderson, S. M. Chignell, A. G. Vorster, R. Lawrence, and P. H. Evangelista. 2017. A survival guide to Landsat preprocessing. Ecology **98**:920-932.

Zimmermann, N. E., T. C. Edwards, C. H. Graham, P. B. Pearman, and J.-C. Svenning. 2010. New trends in species distribution modelling. Ecography **33**:985-989.

Zimmermann, N. E., T. C. Edwards, G. G. Moisen, T. S. Frescino, and J. A. Blackard. 2007. Remote sensing-based predictors improve distribution models of rare, early successional and broadleaf tree species in Utah. Journal of Applied Ecology **44**:1057-1067.

Zurell, D., N. E. Zimmermann, T. Sattler, M. P. Nobis, and B. Schröder. 2016. Effects of functional traits on the prediction accuracy of species richness models. Diversity and Distributions **22**:905-917.

Appendix A

Supporting figures and tables

Appendix A has all supporting figures and tables for chapter 3 and chapter 4.

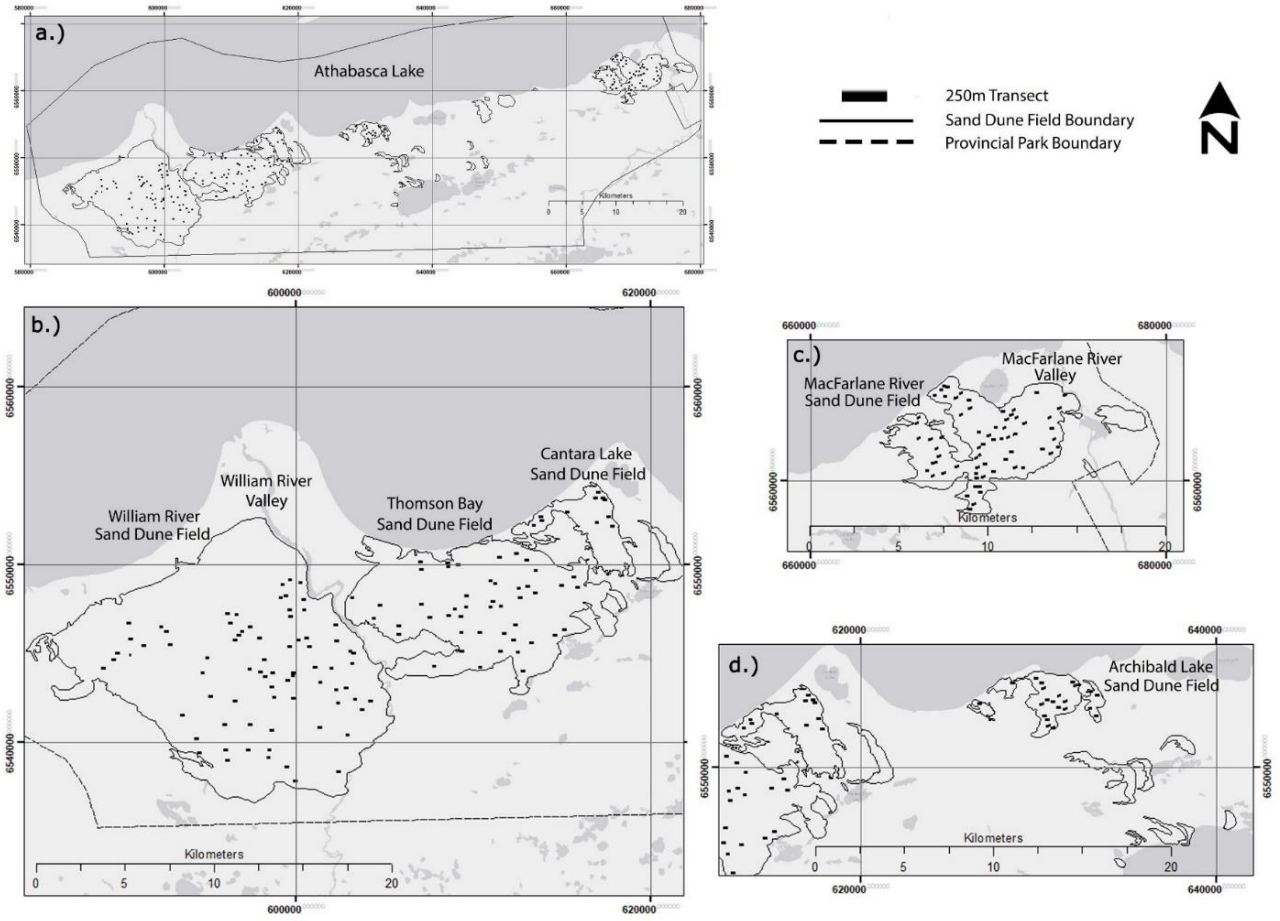


Figure S3.1: Field survey sampling transect plan. a.) Total study site with sampling transects. b.) William River, Thompson Bay, and Cantera Lake sand dune fields close-up. C.) MacFarlane River sand dune field close-up d.) Archibald Lake sand dune field close-up. Each black colour rectangles inside sand dune field indicate 250m transect locations. Transect allocation was stratified random and location by division into 1) ecological strata based on forest stand and dune type polygons derived from air photo interpretation and 2) distance strata where the dune fields were divided into polygons representing the walking distance from likely access points into dune fields. The ecological and distance strata were overlaid to create target polygons and sampling transects were randomly located inside each target polygon. A full description of the sampling methodology available in *Lamb et al.* (2011).

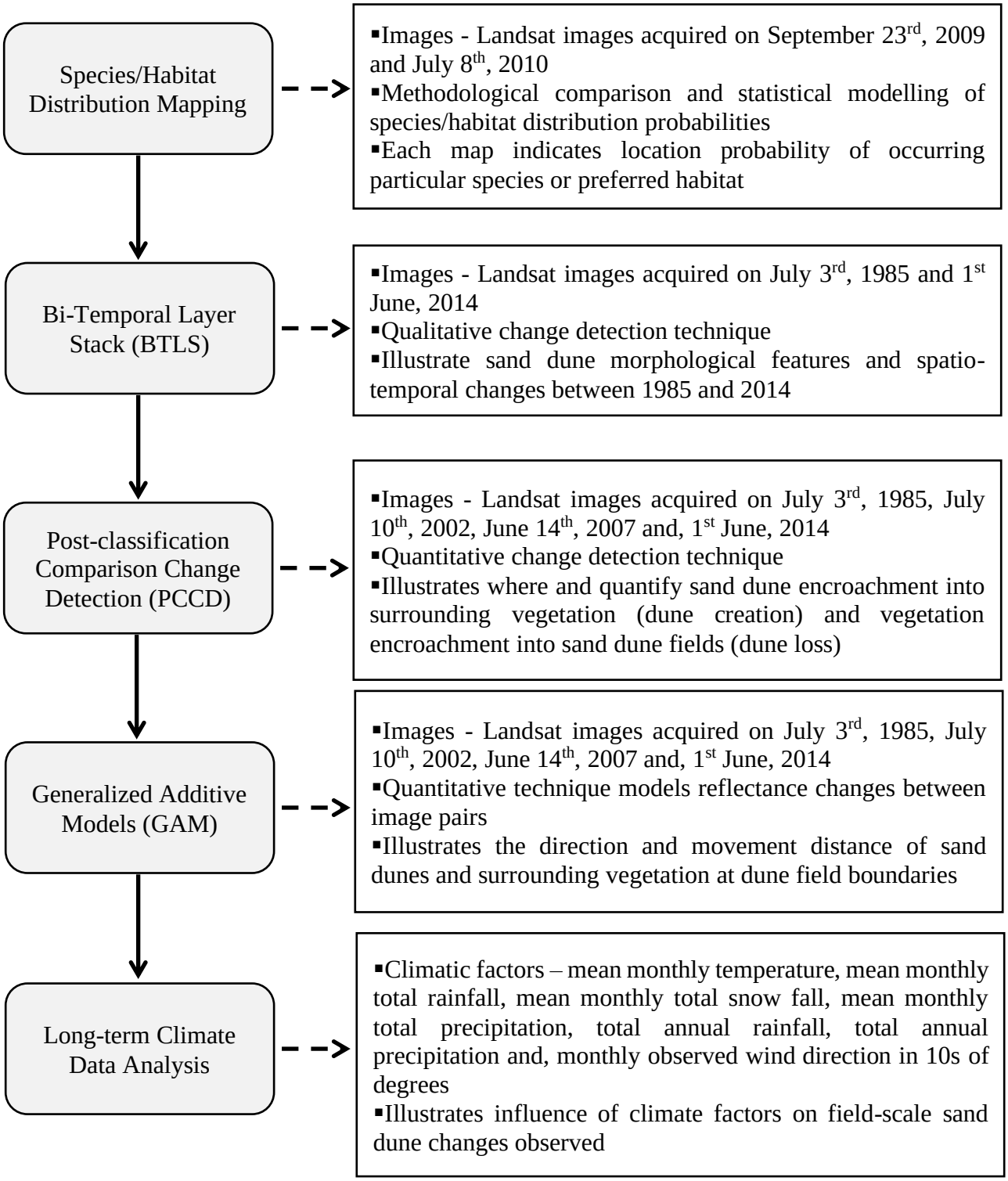


Figure S3.2: Workflow overview. The figure explains detailed steps followed in order and intended uses of each step in the study. The evidence from each step is complementary to each other and support the overall objective of the study.

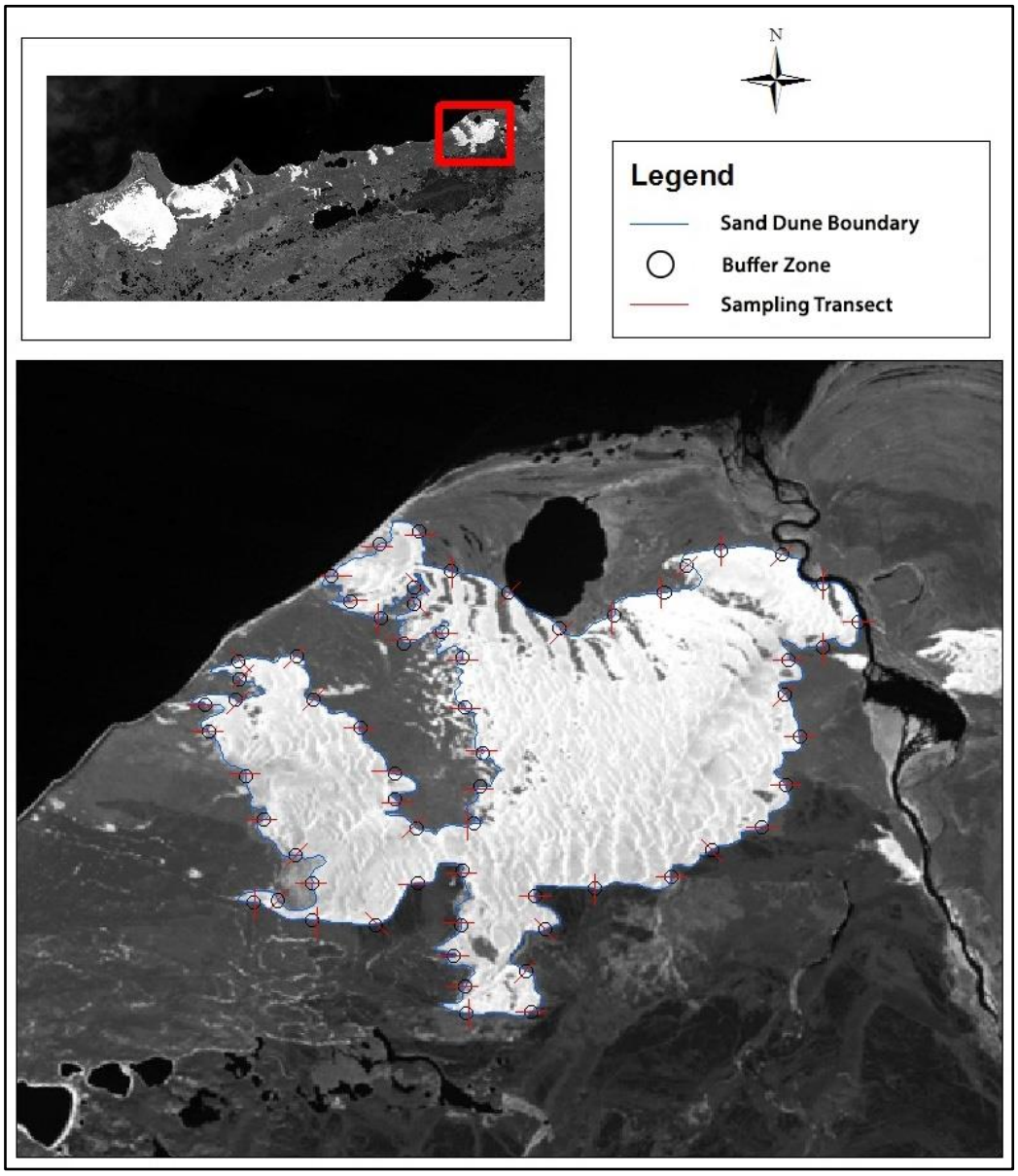


Figure S3.3: Sampling transect map of MacFarlane River dune field. The map contains McFarlane River dune field in Athabasca sand dunes. The blue line around the dune boundary is Dune_Field_Boundary line feature class used to locate sampling points 1 km apart to each other. The black circles are 100 m radius buffer zone (polygon feature class named Sample_Point_100m_Buffer) created at each sample point to locate the transects approximately perpendicular to the dune edge. The red lines at each sample location around the dune field (Sampling_Transect line feature class) were transects used to extract pixel reflectance values from 1985, 2002, 2007 and, 2014 NIR images.

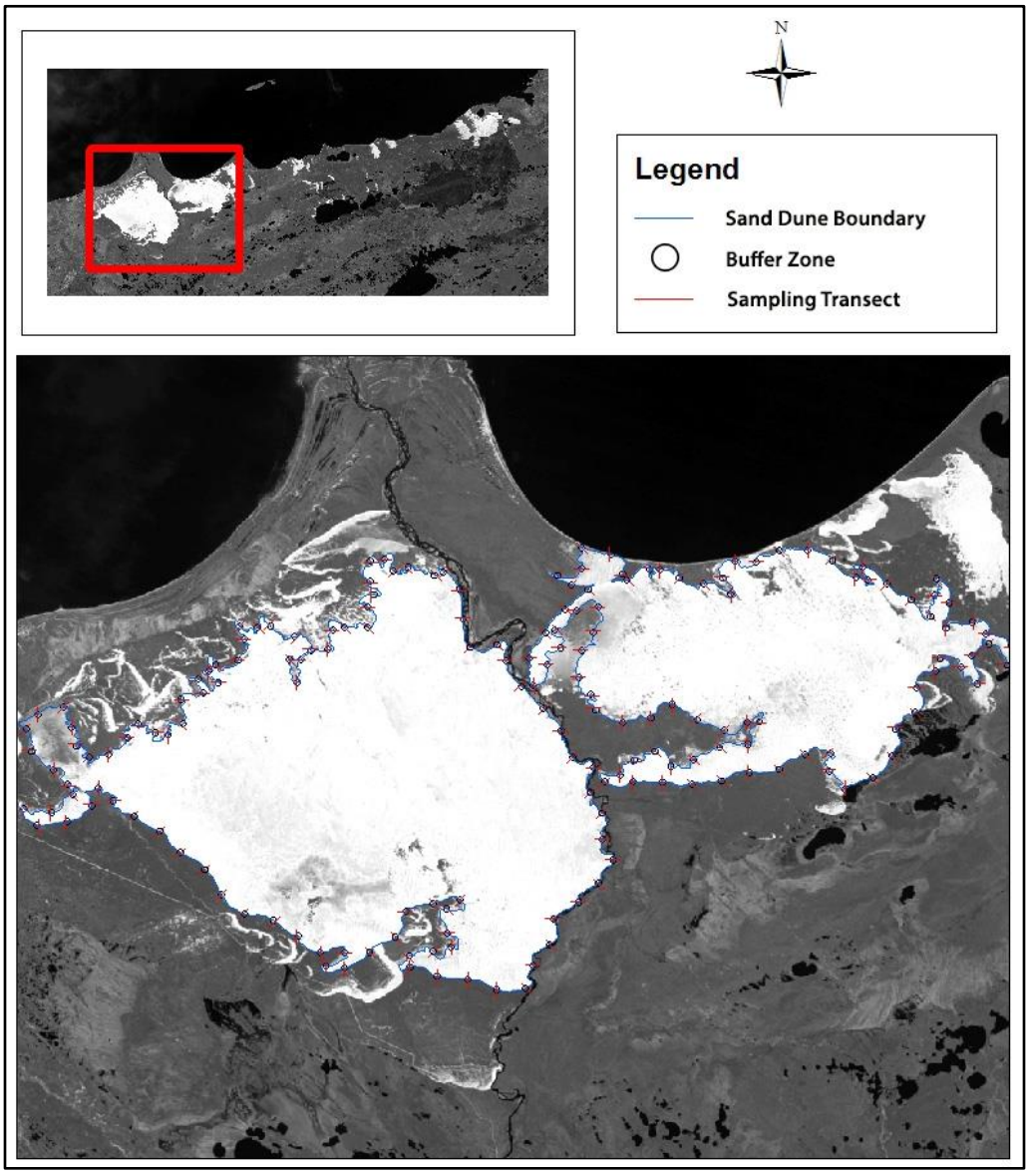


Figure S3.4: Sampling transect map of Thompson Bay, and William River dune fields. The map contains Thompson Bay and William River dune fields in Athabasca sand dunes. The blue line around the dune boundary is Dune_Field_Boundary line feature class used to locate sampling points 1 km apart to each other. The black circles are 100 m radius buffer zone (polygon feature class named Sample_Point_100m_Buffer) created at each sample point to locate the transects approximately perpendicular to the dune edge. The red lines at each sample location around the dune field (Sampling_Transect line feature class) were transects used to extract pixel reflectance values from 1985, 2002, 2007 and, 2014 NIR images.

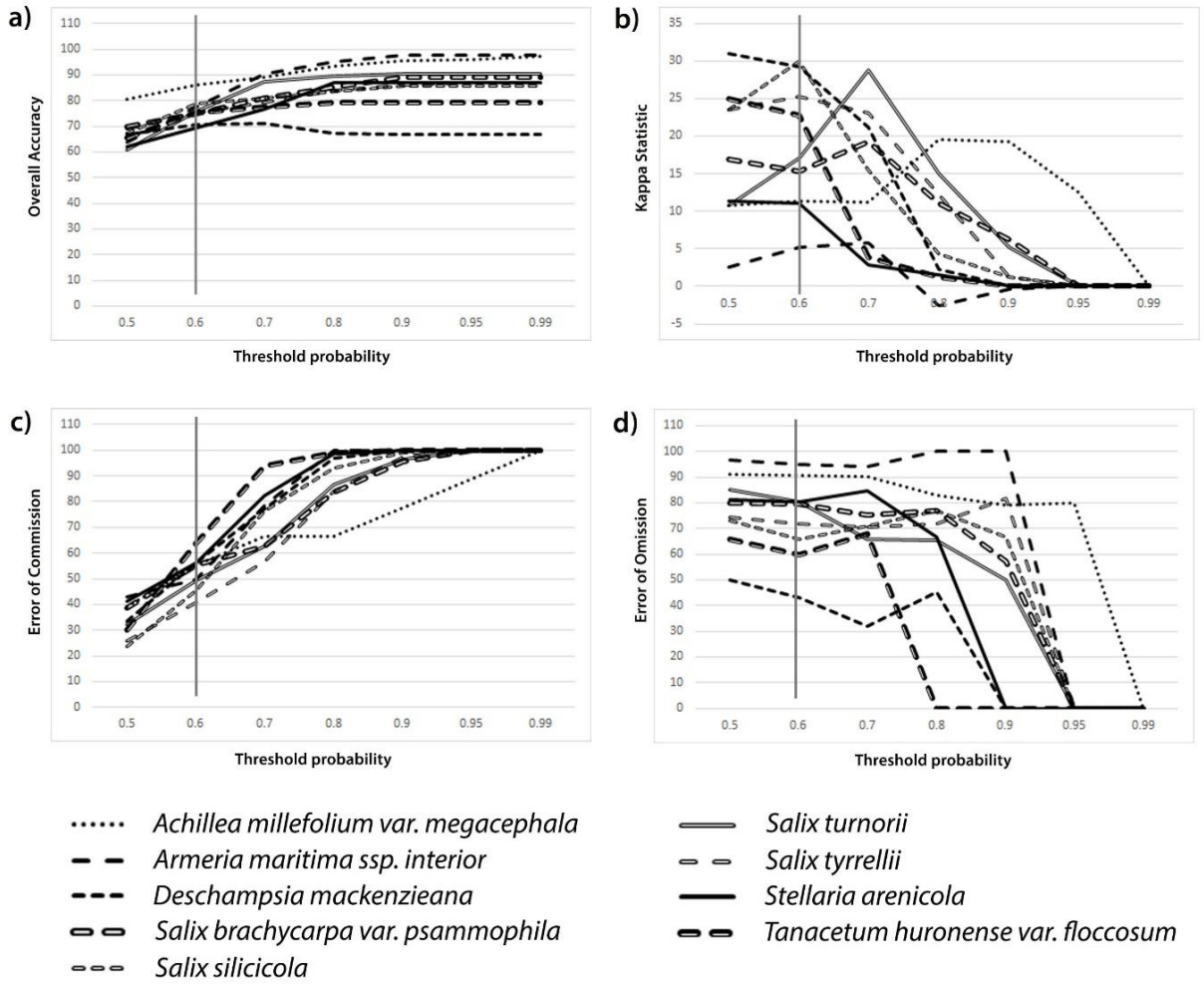


Figure S3.5: Analysis of prediction errors and accuracies based on varying threshold probabilities. a.) Overall accuracy, b.) Kappa statistics, c.) Error of commission, and d.) Error of omission. Each line is a representation of each species and estimate variations are based on various threshold probabilities between 0.5 to 0.99. The line at the 0.6 threshold represents the most optimum point in-comparison to all accuracy measures in consideration.

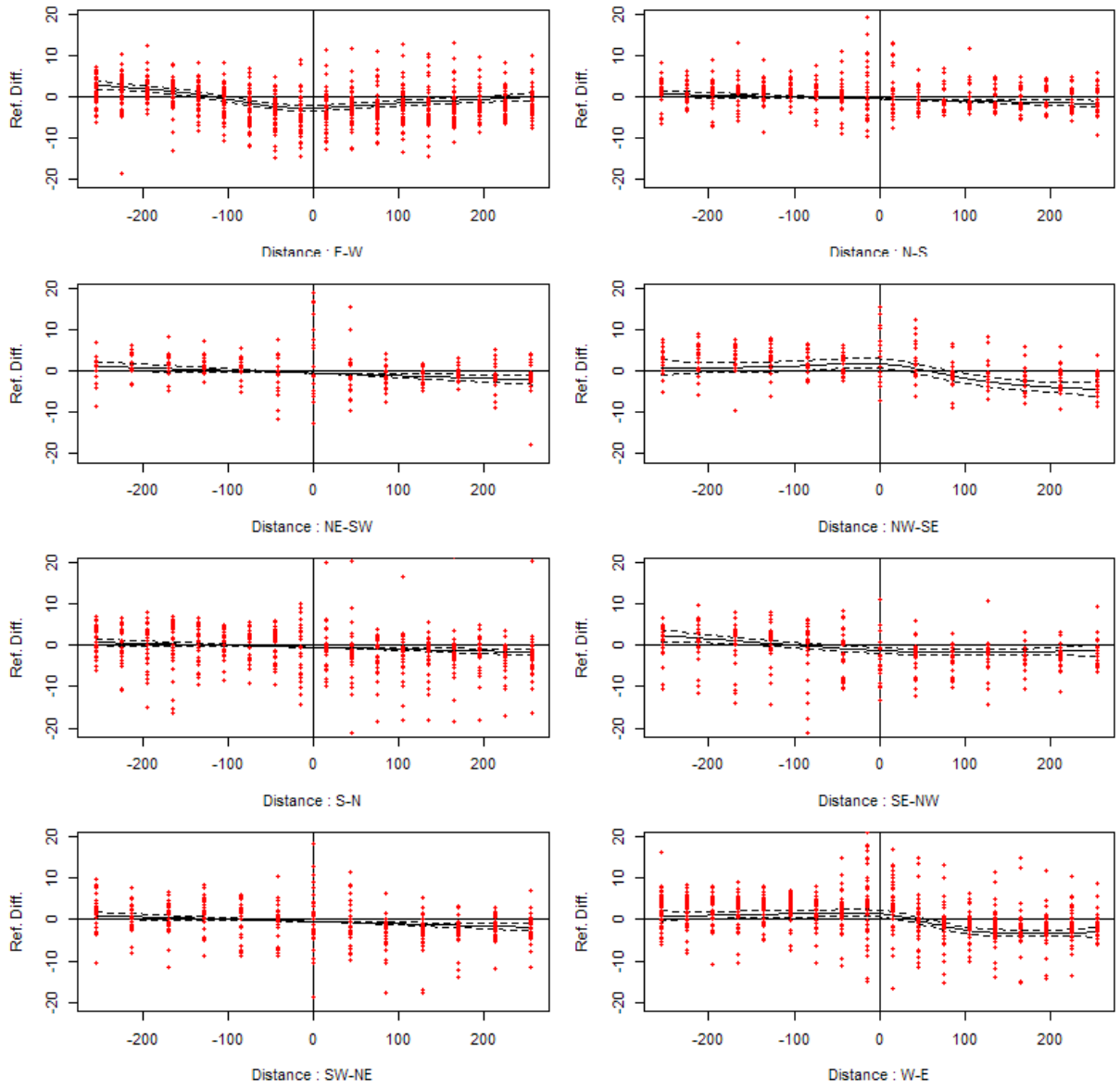


Figure S3.6: Generalized Additive Modelling (GAM) results of directional sand movement analysis from 1985 to 2002. The positive response values indicate sand dune migration to vegetation and the negative response values indicate vegetation encroachment into sand dunes. Response values close to zero means no reflectance differences from 1985 to 2002 observed. The center of the transect was placed approximately at the edge of each sample point on the 1985 base map. The negative distances increase toward the interior of the dunes and the positive distances increase toward the surrounding vegetation. All 8 transect directional categories were separately analyzed.

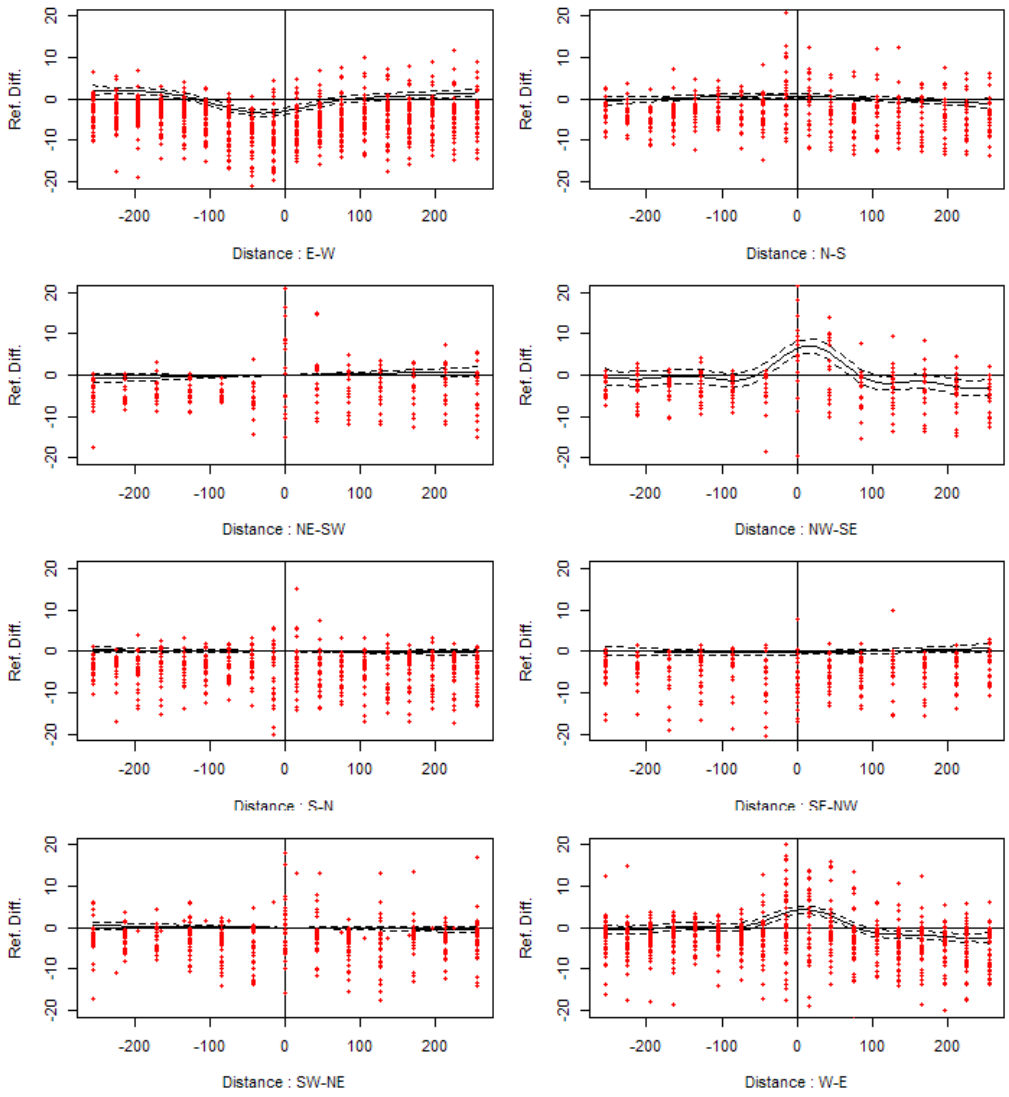


Figure S3.7: Generalized Additive Modelling (GAM) results of directional sand movement analysis from 1985 to 2007. The positive response values indicate sand dune migration to vegetation and the negative response values indicate vegetation encroachment into sand dunes. Response values close to zero means no reflectance differences from 1985 to 2007 observed. The center of the transect was placed approximately at the edge of each sample point on the 1985 base map. The negative distances increase toward the interior of the dunes and the positive distances increase toward the surrounding vegetation. All 8 transect directional categories were separately analyzed.

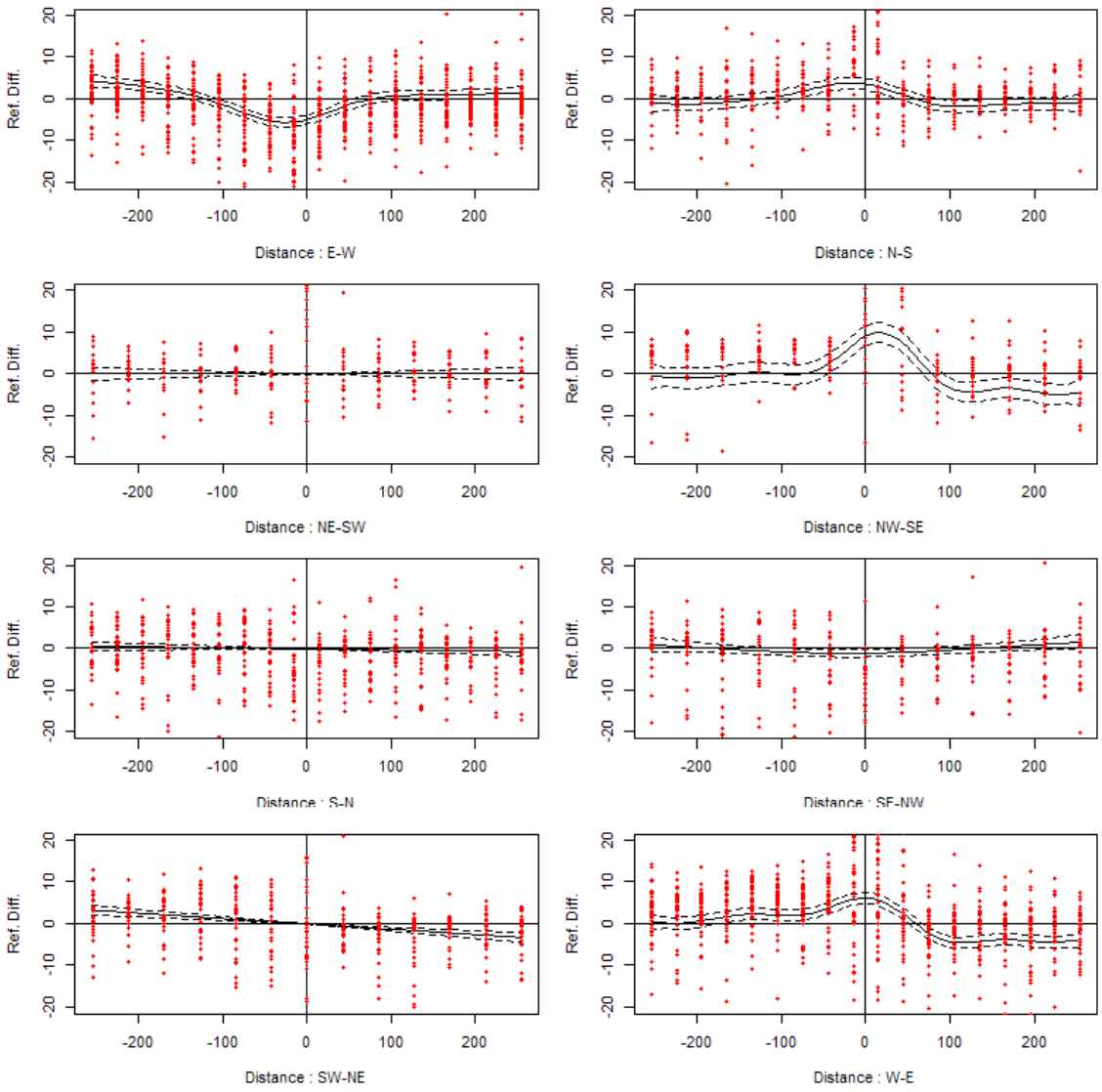


Figure S3.8: Generalized Additive Modelling (GAM) results of directional sand movement analysis from 1985 to 2014. The positive response values indicate sand dune migration to vegetation and the negative response values indicate vegetation encroachment into sand dunes. Response values close to zero means no reflectance differences from 1985 to 2014 observed. The center of the transect was placed approximately at the edge of each sample point on the 1985 base map. The negative distances increase toward the interior of the dunes and the positive distances increase toward the surrounding vegetation. All 8 transect directional categories were separately analyzed.

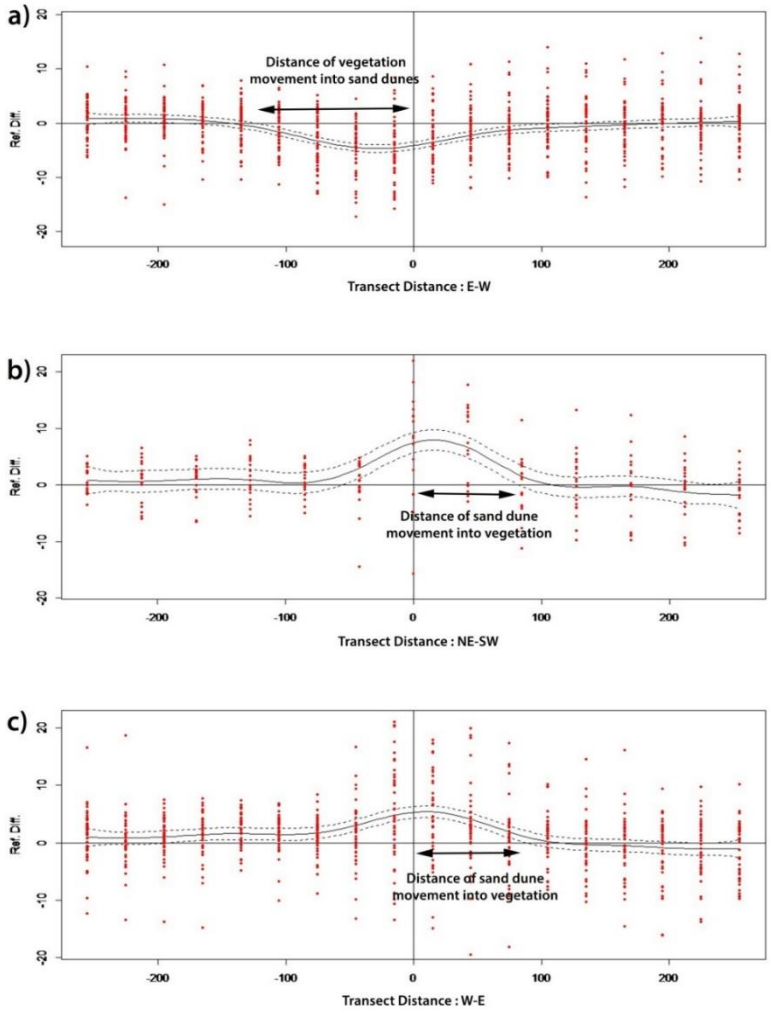


Figure S3.9: Illustration of sand dune and vegetation movement distance calculation using Generalized Additive Models (GAM). a) Illustrates the direction (E-W) where negative significant reflectance difference was observed. The distance was calculated from zero (edge of the dune field) to a point where upper confidence limit crosses the main X-axis. The direction of the distance was into the sand dunes. b and c) Illustrates the direction (NW-SE and W-E) where positive significant reflectance difference was observed. The distance was calculated from zero (edge of the dune field) to a point where lower confidence limit crosses the main X-axis. The direction of the distance was into the vegetation.

Table S3.1: Habitat types identified during fieldwork in the Athabasca sand dunes following Lamb et al. (2011), and habitat types used in the present study.

The first column corresponds to the Habitat types identified during the field surveys. The last column explains the observed characteristics of the habitat.

Field Type	Name	Description
WIDS	Wet inter-dune slack	Wet inter-dune slack. A level or nearly level landscape with a high groundwater table and moist soils. Open water is occasionally present. May have sandy substrate or more or less extensive herbaceous or bryophyte ground cover.
SIDS	Saline inter-dune slack	Saline inter-dune slack. As for WIDS but with evidence of salt deposits on the soil surface.
LSDN	Low-slope gradient dune	Dry low-slope gradient dune. Dominant substrate is open sand with slopes generally less than 15-20°. Relatively level areas of open sand between dunes without evidence of a high water table were included in this category.
HSDN	High-slope gradient dune	Dry high-slope gradient dune. Dominant substrate is open sand with slopes generally greater than 15-20°.
GRPV	Gravel pavement	Gravel pavement. Dominant surface cover is rocks or pebbles lying on a sandy substrate.
LICH	Lichen-crowberry heaths	Lichen-crowberry heaths. Dry areas with well developed layers of lichens, bryophytes,and low-growing ericaceous shrubs over the soil surface, but without extensive tall shrub or tree cover.
WOOD	Woodland	Extensive woody vegetation (generally jackpine forest or birch scrub). Substrates between trees generally similar to LICH.

Table S3.2: Analysis of estimate uncertainty based on varying threshold probabilities.

Species abbreviations refer to *Achillea millefolium* var. megacephala (ACHMIL), *Armeria maritima* ssp. Interior (ARMMAR), *Deschampsia mackenzieana* (DESMAC), *Salix brachycarpa* var. psammophila (SALBRA), *Salix silicicola* (SALSIL), *Salix turnorii* (SALTUR), *Salix tyrrellii* (SALTYR), *Stellaria arenicola* (STEARE), and *Tanacetum huronense* var. floccosum (TANHUR). There are three estimates calculated for each threshold a.) Total occupied habitat extent estimate (km^2) b.) Affected (stabilized) habitat extent between 1985 to 2014 (km^2) and c.) Percent of estimated habitat influenced by sand dune stabilization.

Threshold		ACHMIL	ARMMAR	DESMAC	SALBRA	SALSIL	SALTUR	SALTYR	STEARE	TANHUR
0.5	A (km^2)	58.35	82.45	154.10	119.11	105.26	108.02	137.62	126.77	137.17
	B (km^2)	28.51	13.55	12.53	37.00	30.65	30.29	41.47	42.83	39.04
	C (%)	48.87	16.43	8.13	31.06	29.12	28.04	30.13	33.79	28.46
0.6	A (km^2)	38.92	48.82	89.52	80.65	59.45	66.27	100.38	80.30	57.92
	B (km^2)	20.85	7.96	8.22	30.48	22.13	22.68	35.72	35.50	21.75
	C (%)	53.57	16.30	9.18	37.79	37.23	34.22	35.58	44.20	37.55
0.7	A (km^2)	26.99	21.31	38.62	47.24	27.83	32.64	60.31	37.18	13.83
	B (km^2)	15.19	3.60	4.45	21.96	13.58	14.20	26.78	21.75	6.16
	C (%)	56.30	16.88	11.52	46.49	48.78	43.51	44.40	58.50	44.52
0.8	A (km^2)	17.12	6.35	9.52	20.82	8.70	11.54	23.52	0.91	1.18
	B (km^2)	9.86	1.22	1.23	11.55	5.19	5.72	13.85	0.44	0.45
	C (%)	57.61	19.14	12.92	55.46	59.63	49.57	58.88	48.27	38.53
0.9	A (km^2)	8.22	0.74	0.63	3.57	0.70	1.48	2.88	0.00	0.02
	B (km^2)	4.77	0.14	0.02	2.21	0.42	0.68	2.04	0.00	0.01
	C (%)	57.99	19.19	2.85	61.68	60.36	45.95	70.70	0.00	26.92
0.99	A (km^2)	0.76	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00
	B (km^2)	0.45	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	C (%)	58.98	22.22	0.00	22.22	0.00	0.00	33.33	0.00	0.00

Table S3.3: Iterative Self-Organizing Data (ISODATA) unsupervised classification statistics.

The table illustrates ISODATA unsupervised classification results of 1985, 2002, 2007 and, 2014 images. The column I.C stands for initial classification classes obtained from ISODATA procedure and F.C stands for final combined classes. The mean, standard deviation and, number of pixels of each initial class are provided in the table. The similar classes were merged to obtain three distinct land cover classes; water (W), vegetation (V) and open sand (S) of the area. The column final class (F.C) illustrates how each initial classes were combined to three distinct land cover classes.

	1985 Image Classification				**2002 Image Classification**				**2007 Image Classification**				**2014 Image Classification**			
I.C	**Mean**	**Std. Dev**	**Pixels**	**F.C**	**Mean**	**Std. Dev**	**Pixels**	**F.C**	**Mean**	**Std. Dev**	**Pixels**	**F.C**	**Mean**	**Std. Dev**	**Pixels**	**F.C**
1	0.50	0.37	1674453	W	4.91	0.65	2265373	W	0.25	0.46	2299709	W	17.03	1.63	493768	W
2	7.50	1.83	481064	V	15.96	2.66	126176	W	7.40	1.90	559803	V	25.32	2.30	443885	W
3	16.07	3.10	311761	V	29.30	2.89	582306	V	16.86	1.53	690182	V	30.89	1.36	1546267	V
4	9.21	1.54	783163	V	22.13	1.66	988666	V	22.61	2.27	382814	V	35.64	2.18	799059	V
5	14.24	1.75	792580	V	23.14	1.39	822118	V	11.33	1.31	787002	V	21.78	3.46	166355	W
6	1.96	0.94	561619	W	7.87	3.00	46333	W	16.63	1.87	463133	V	37.34	2.77	327805	V
7	18.93	1.75	543956	V	25.43	1.34	433048	V	18.97	2.05	193032	V	29.94	2.02	137949	W
8	23.05	2.30	163360	V	27.61	1.80	198019	V	22.61	1.88	103757	V	49.33	3.92	49399	V
9	3.96	1.11	156622	W	32.10	2.57	39546	V	25.61	2.15	37670	V	34.90	1.54	213595	W
10	28.02	3.20	38441	V	37.95	2.44	23438	V	30.92	2.25	19897	V	61.83	3.05	56313	V
11	34.44	3.03	19851	V	43.55	2.29	18590	V	36.49	2.07	13966	V	38.43	1.15	229913	W
12	40.80	2.95	14866	V	48.68	2.16	16599	V	41.19	1.82	13868	V	69.18	1.90	113431	S
13	45.65	2.88	15491	V	52.93	1.93	18675	V	44.29	1.69	20694	V	41.58	0.98	375696	W
14	51.15	2.46	23821	S	56.23	1.69	27447	S	46.84	1.41	34134	S	74.13	1.72	120164	S
15	55.01	2.33	33206	S	59.16	1.27	46165	S	49.15	1.16	49663	S	44.44	1.05	309318	W
16	58.59	2.02	46443	S	61.72	0.99	71301	S	51.24	0.96	59458	S	47.31	0.98	297896	W
17	60.87	1.55	56247	S	64.46	0.96	77892	S	52.94	0.85	61049	S	50.10	1.29	169741	W
18	63.73	1.19	62219	S	67.01	1.21	48862	S	54.58	0.81	45368	S				
19	66.22	1.12	49378	S					56.88	1.16	15355	S				
20	68.91	1.38	22013	S												

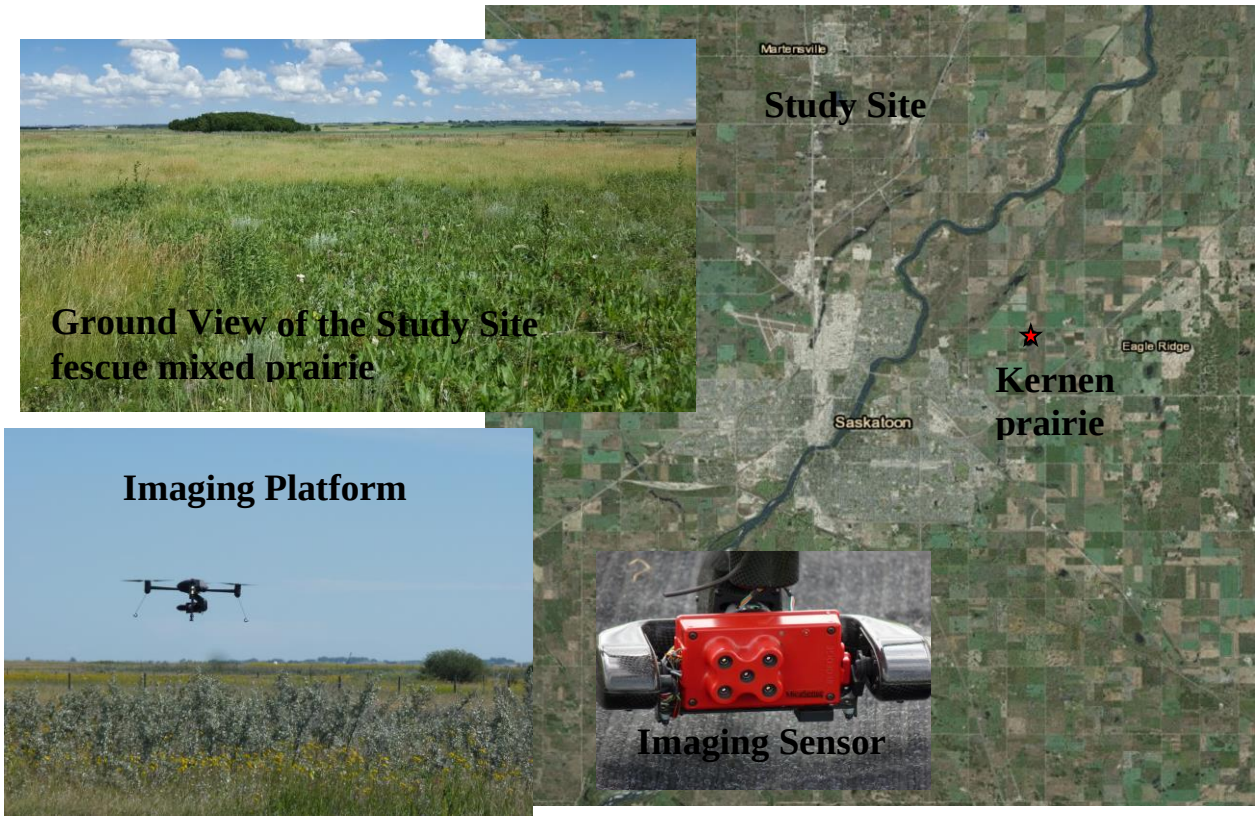


Figure S4.1: Study site location, Ground view, Imaging Platform Draganfly X4P, and Imaging Sensor MicaSense RedEdge multispectral (5 bands).

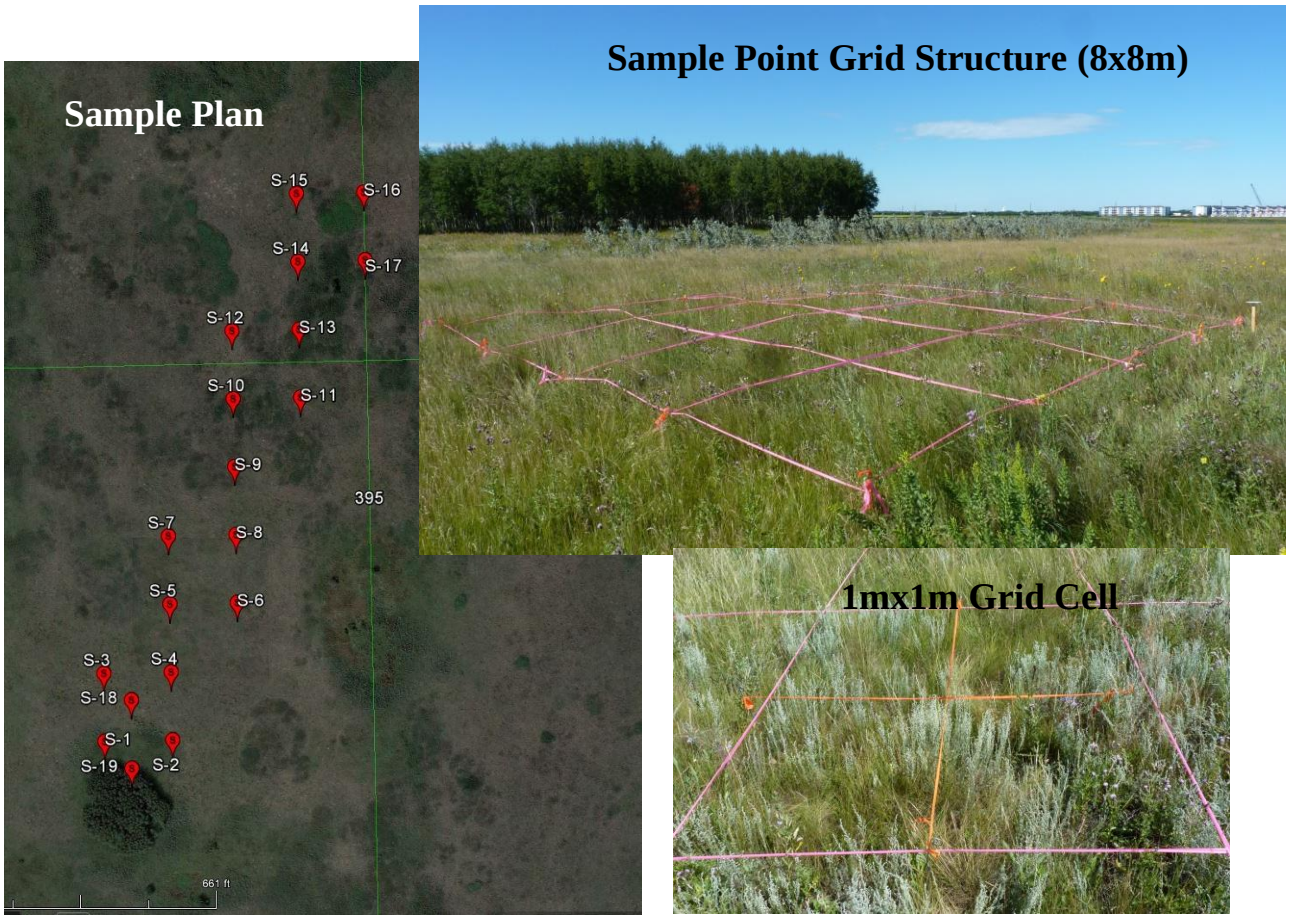


Figure S4.2: Plant sampling survey plan, sample point grid structure, and grid cell structure.

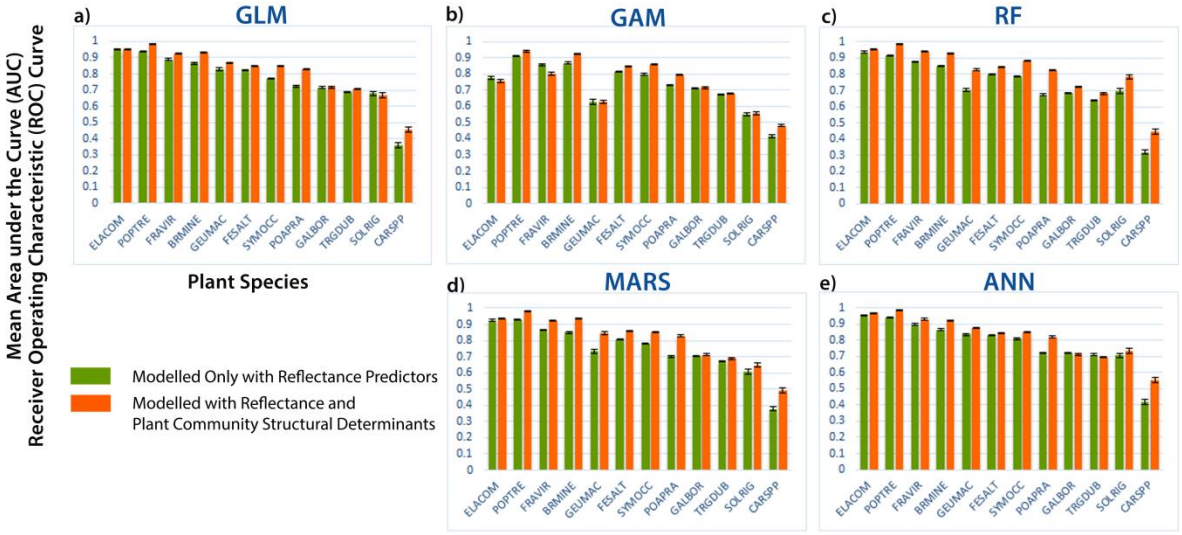


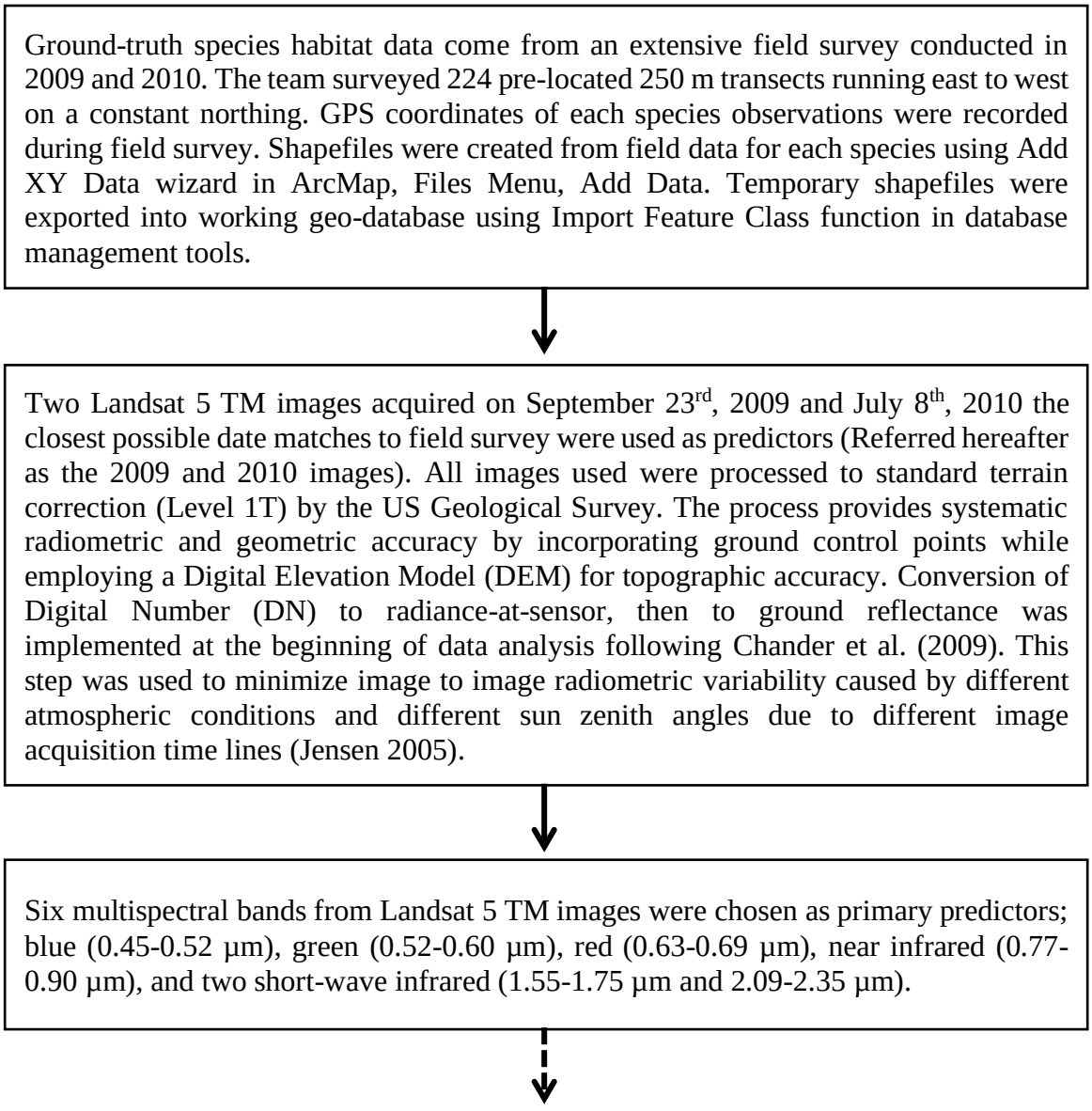
Figure S4.3: Receiver Operating Characteristic (ROC) Curve Comparison with and without Plant Community Structural Determinants/Predictors (Average height of the plant community and average litter thickness). The Y-axis is the mean Area Under the Curve (AUC) of the ROC value for each species and each modelling algorithm. Each bar for a species represents the calculated mean AUC of hundred iterations of an algorithm. Each species has AUC comparison between modelling only with reflectance predictors (GREEN) and modelling together with reflectance and plant community structural determinants (ORANGE). Species abbreviations refer to *Elaeagnus commutate* (ELACOM), *Populus tremuloides* (POPTRE), *Fragaria virginiana* (FRAVIR), *Bromus inermis* (BRMINE), *Geum macrophyllum* (GEUMAC), *Festuca altaica* (FESALT), *Symphoricarpos occidentalis* (SYMOCC), *Poa pratensis* (POAPRA), *Galium boreale* (GALBOR), *Tragopogon dubius* (TRGDUB), *Solidago rigida* (SOLRIG), *and Carex spp* (CARSPP). Modelling techniques include Generalized Linear Models (GLM), Generalized Additive Models (GAM), Random Forest (RF), Multivariate Adaptive Regression Splines (MARS), and Artificial neural networks (ANN).

Appendix B

Detailed methods for chapter 3

B.1 Habitat/species distribution modelling

Habitat/species modelling was implemented in R 3.1.2 software and ArcGIS 10.3 software environments. Landsat images acquired on September 23rd, 2009 and July 8th, 2010 were primary predictors of the modelling process. Nine species were targeted; that include *Achillea millefolium* var. *megacephala* and *Tanacetum huronense* var. *floccosum* (both Asteraceae), *Armeria maritima* ssp. *interior* (Plumbaginaceae), *Deschampsia mackenzieana* (Poaceae), *Salix brachycarpa* var. *psammophila*, *Salix silicicola*, *Salix turnorii*, and *Salix tyrrellii* (Salicaceae), and *Stellaria arenicola* (Caryophyllaceae).



Reflectance values were extracted from all six spectral bands along each transect that corresponds to species presence or absence. The extraction was performed using Sample tool in ArcGIS, Arc Tool Box, Spatial Analyst Tools, Extraction. Species presence or absence locations feature classes were used as the input location raster or point feature to identify pixel locations.

↓

The input data set was composed of x and y geographic coordinates of the position, presence or absence of each species, and reflectance values from all six multispectral bands of the location. Species abbreviations used in the data set were *ACHMIL - Achillea millefolium, ARMMAR - Armeria maritima, DESMAC - Deschampsia mackenzieana, SALBRA - Salix brachycarpa, SALSIL - Salix silicicola, SALTUR - Salix turnorii, SALTYR - Salix tyrrellii, STEARE - Stellaria arenicola,* and *TANHUR - Tanacetum huronense.* Six multispectral bands were BAND 1 - blue (0.45-0.52 µm), BAND 2 - green (0.52-0.60 µm), BAND 3 - red (0.63-0.69 µm), BAND 4 - near infrared (0.77-0.90 µm), BAND 5 - short-wave infrared (1.55-1.75 µm), and BAND 6 – short-wave infrared (2.09-2.35 µm).

↓

The BIOMOD2 package was used in R 3.1.2 software version for model evaluation and training. Five different modelling algorithms were evaluated; generalized linear models (GLM), generalized additive models (GAM), multivariate adaptive regression splines (MARS), classification and regression trees (CART), and artificial neural networks (ANN). Each algorithm and species combination used 80% data split between trials and ran through 1000 iterations. The study used threshold-independent measures of accuracy; "area under the curve" (AUC) of the receiver-operating characteristic (ROC) plot to select the most suitable method for modeling target plant species (Figure 2).

↓

The best modelling algorithm for a species was selected comparing man AUC of 1000 iterations. Welch's One-way Analysis of Variances procedure and Games-Howell pair-wise mean comparison was used to illustrate differences and/or similarities of mean AUC among modeling algorithms for a given species (Table 1). The model formula with the highest AUC of selected algorithm was used for final habitat distribution modeling in ArcGIS 10.3 software environment.

⇣

The final model formula for a given species was imported into ArcGIS geoprocessing model builder for final probability map generation. The model builder was used to graphically draft development process and the model formula for a given species. The Raster Calculator tool, in ArcGIS, Arc Tool Box, Spatial Analysis Tools, Map Algebra was used to mathematically draft the formula. The final result of the raster processing was saved to the working geo-database. 2009 and 2010 data were separately modelled and each location probabilities were averaged using Raster Calculator tool, in ArcGIS, Arc Tool Box, Spatial Analysis Tools, Map Algebra to generate final probability surface for each species.

↓

Final probability surface of each species was reclassified into 100 classes using Equal Interval Classification in Table of Content Window, Layers, Layer Properties, Classified dialog box. Color ramp was assigned to probability classes; green – low, yellow – mid, and red – high for visual demarcation of geographic probability distribution (Figure 3).

↓

Estimation of occupied/suitable habitat for each species was assessed across a range of thresholds (0.5, 0.6, 0.7, 0.8, 0.9, and 0.99) to decide the most appropriate cut-off for Athabasca endemics (Figure 4 a and b). Binary (yes/no) layers of each cut-off were produced using Raster Calculator tool, in ArcGIS, Arc Tool Box, Spatial Analysis Tools, Map Algebra. Then the occupied/suitable area was estimated multiplying pixel count with area of a pixel ($900m^2$).

↓

Prediction accuracies were evaluated using 30% of ground truth data held back from the model training process. A series of threshold probabilities (0.5, 0.6, 0.7, 0.8, 0.9, 0.95 and 0.99) were used to illustrate changes in error of commission, error of omission, overall accuracy, and kappa statistic. All accuracy measures were calculated based on the confusion matrix of each species for a given threshold.

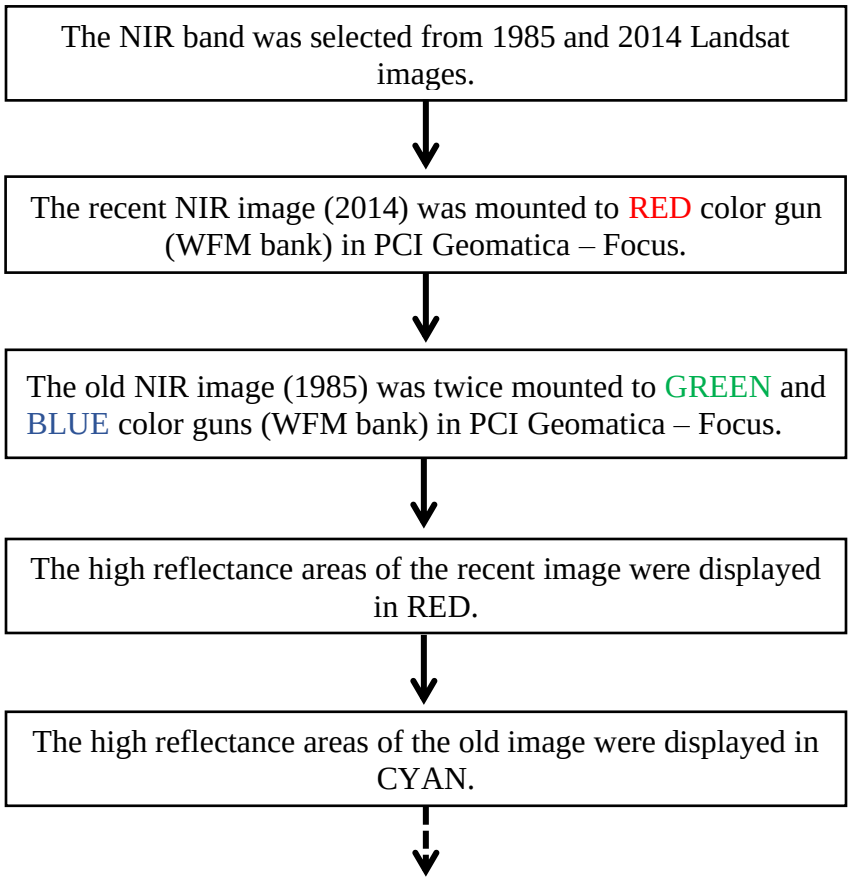
⇣

Classification overlay analysis was used to estimate sand dune stabilization influence on each species habitat extent estimates. Initially, each species prediction probability surface was converted to a binary (yes/no) response surface. A series of threshold probabilities (0.5, 0.6, 0.7, 0.8, 0.9, 0.95 and 0.99) were used to illustrate variations in binary estimates on the basis of different thresholds (Figure 4 c and d). The estimation was done using Raster Calculator tool, in ArcGIS, Arc Tool Box, Spatial Analysis Tools, Map Algebra.

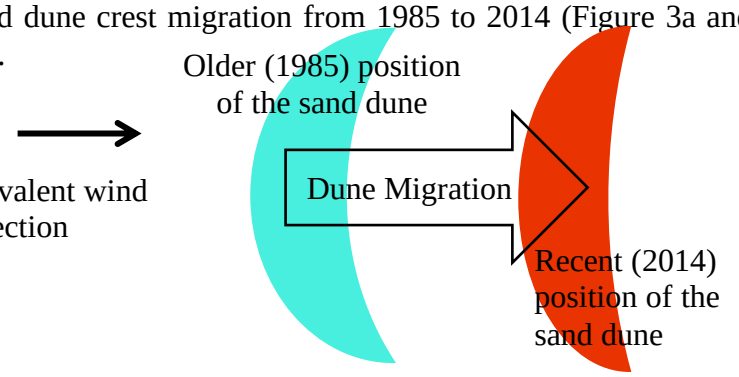
Overlay analysis was performed between binary predictions of varying thresholds and stabilized sand dune area estimate between 1985 to 2014. Prior to perform overlay analysis, all binary "No" classes were set to null using Set Null tool, in ArcGIS, Arc Tool Box, Spatial Analysis Tools, Conditional. The final overlay results were obtained multiplying two raster layers in comparison using Raster Calculator tool, in ArcGIS, Arc Tool Box, Spatial Analysis Tools, Map Algebra. The result of the overlay analysis had two classes; 1. Null class – No overlapping classes and 2. Influenced habitat – estimated habitat of a species overlaps with stabilized dune area. Then, the habitat area influenced were estimated multiplying pixel count by the area of a pixel ($900m^2$).

B.2 Bi-temporal layer stack (BTLS)

The BTLS is an alteration of the Write Function Memory Insertion (WFMI) technique. The method is well established qualitative procedure to visually examine land use and land cover temporal changes. The technique basically requires bi-temporal or multi-temporal images of the same band (wavelength) and the same geographical coverage. The study used bi-temporal images as it is the most appropriate for the context.

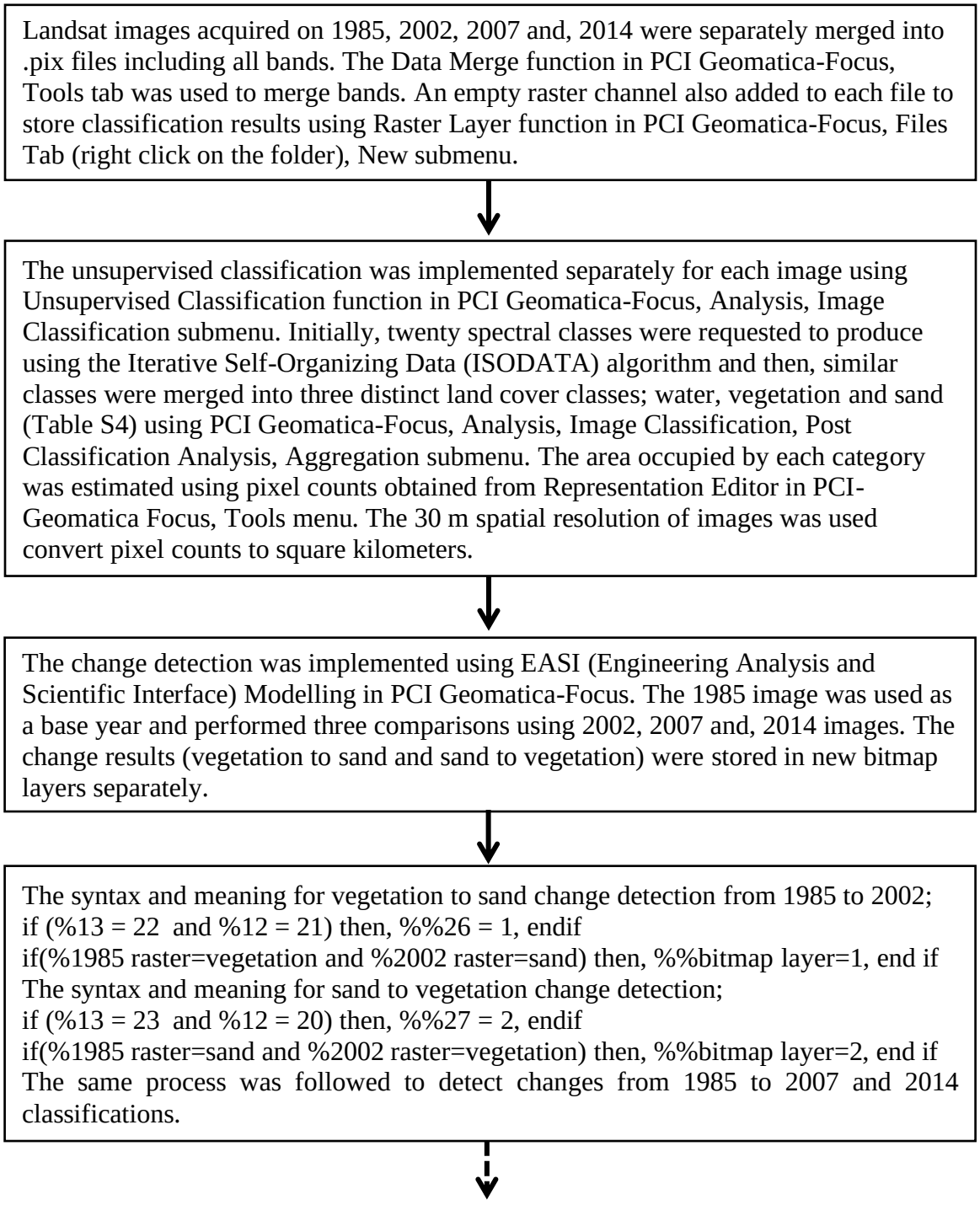


The adjacent position of RED and CYAN color areas indicates sand dune crest migration from 1985 to 2014 (Figure 3a and 3b).



B.3 Post-classification comparison change detection (PCCD)

The PCCD procedure was implemented to understand how sand dune encroachment occurs into surrounding vegetation and how vegetation encroachment occurs into sand dune fields at the edge of McFralane river, William river and Thompson bay dune field areas.

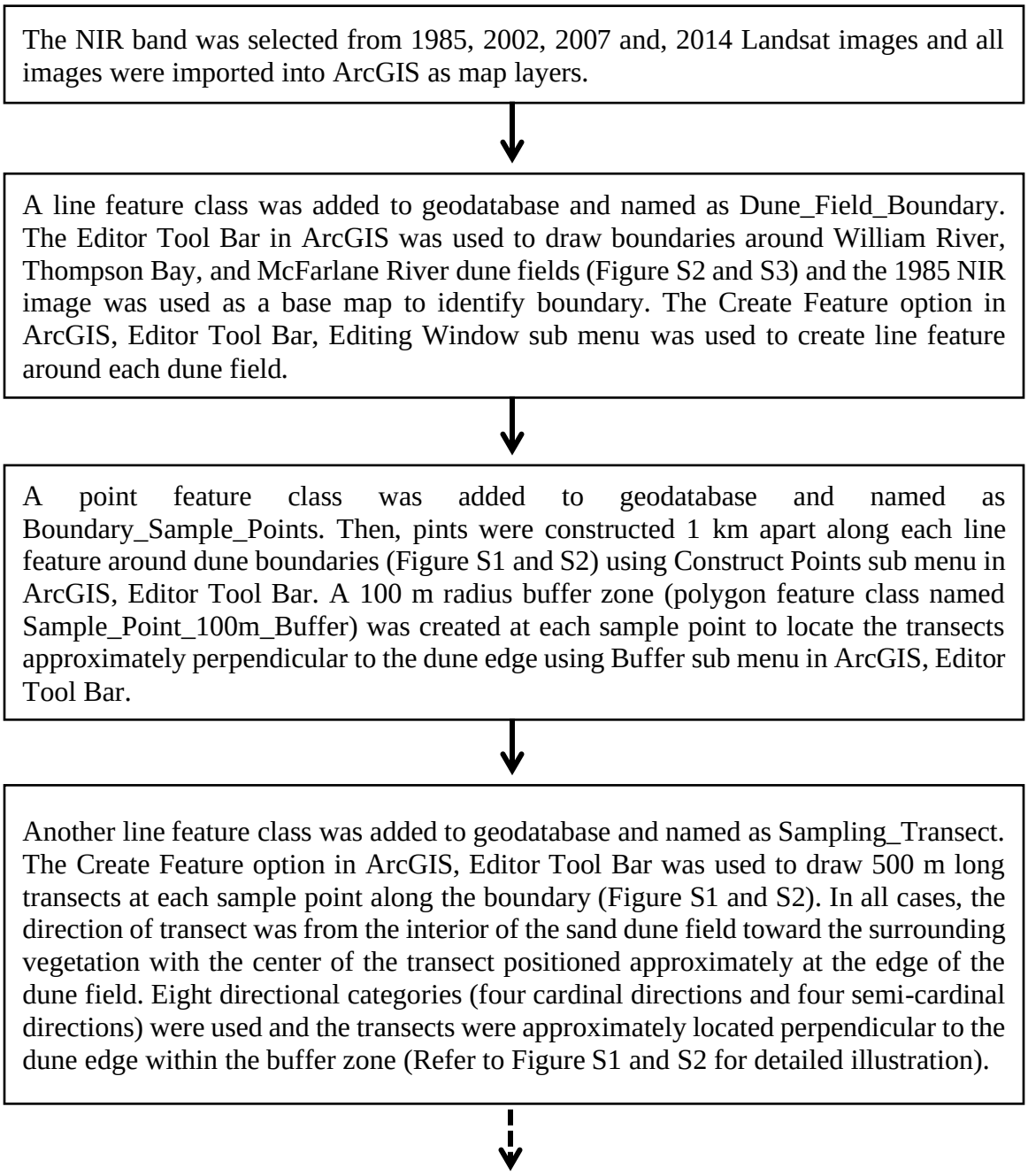


The area change (vegetation to sand and sand to vegetation) was calculated using pixel counts of each bitmap layer produced. The pixel counts were obtained from Representation Editor in PCI-Geomatica Focus, Tools menu. The number of pixels was converted to square kilometers using 30 m spatial resolution of images. The annual rate of the change was calculated based on the time interval of the image pairs.

The final Post-classification Comparison Change Detection (PCCD) map was produced overlaying change results from 1985 to 2014 on the base year natural color image to graphically illustrate overall sand dune and woody vegetation dynamics in the area (Figure 4a and 4b).

B.4 Generalized additive modelling (GAM) approach to estimate directions and movement distances of sand dune and vegetation at dune boundaries

An important goal of this study was to quantify the distance and directional movement of dune fields and vegetation encroachment over time.



Another point feature class was added to geodatabase and named as Transect_Sample_Points. Then, pints were constructed 30 m apart for cardinal directions and 42.43 m apart for semi-cardinal directions along each transect around dune boundaries using Construct Points sub menu in ArcGIS, Editor Tool Bar.

The reflectance value of each pixel underneath each transect were extracted from all NIR images (1985, 2002, 2007 and, 2014) using Sample Tool in ArcGIS, Arc Tool Box, Spatial Analysis Tools, Extraction sub menu. The Transect_Sample_Points feature class was used as the input location raster or point feature to identify pixel locations.

The reflectance differences of pixels inline from each recent year (2002, 2007 and 2014) to base year 1985 were calculated and a correction was applied to minimize overall reflectance deviations of both images. The average of pixel reflectance difference at both ends of the transects were calculated as a correction factor (Calculation steps available in R Script_GAM_Athabasca Sand Movement Analysis.R) under the assumption that the transect end pixel reflectance difference should be close to zero (pseudo-invariant feature) as transect ends are always 250 m inside the sand dune or vegetation. Final reflectance difference values were obtained after deducting correction factor from all reflectance difference values of each image pair.

The difference of reflectance from the recent year (2002, 2007 and 2014) to 1985 was modeled as a function of distance along transects (Calculation steps available in R Script_GAM_Athabasca Sand Movement Analysis.R) using Generalized Additive Models (GAM). Positive reflectance differences were sand dune migration into the surrounding vegetation and negative differences were vegetation encroached into the sand dune fields. The center of the pixels underneath each transact were 30 m apart for the transects on the cardinal directions and 42.43 m apart for the semi-cardinal directions. The center of the transact was marked as zero distance with negative distances into the sand dunes, and positive into the surrounding vegetation. Each directional category (four cardinal directions and four semi-cardinal directions) was separately modelled and analyzed to identify directional movement of sand dunes and vegetation (Figure S4, S5, S6 and Table 6).

All 8 GAMs by each directional category were assessed on the basis of $p \leq 0.05$ to identify significant reflectance differences relationship, in relation to transect distance. The significant reflectance difference relationships across all three comparisons (2002, 2007 and, 2014 with 1985 image) were used to estimate distances of negative or positive reflectance relationship observed.

Each individual significant GAM relationship was separately analyzed to calculate distance where positive or negative reflectance difference observed. A sequence of 1000 numbers was generated between -250 and +250 range, corresponding to the distance of transect used in the study. Predictions of the fitted model and confidence interval were calculated based on the distance values generated and a graph was drawn to estimate distances (calculation steps available in File S3: R script file). The distance where observed negative reflectance difference was calculated from zero distance (edge of the dune field) to upper confidence limit crosses the main X-axis. The direction of the distance was into the sand dunes (Figure S7a). The distance where observed positive reflectance difference was calculated from zero distance (edge of the dune field) to lower confidence limit crosses the main X-axis. The direction of the distance was into the vegetation (Figure S3b and S3c).