Predicting Eurasian Watermilfoil and Curly-leaf Pondweed Presence-Absence in Adirondack Lakes Using Geographically Weighted Logistic Regression

by

Jelena Grbic
Bachelor of Science, McGill University, 2016

A Major Research Paper

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Master of Spatial Analysis

in the program of

Spatial Analysis

Toronto, Ontario, Canada, 2017

Author's Declaration for the Electronic Submission of a MRP

I hereby declare that I am the sole author of this MRP. This is a true copy of the MRP, including any required final revisions. I authorize Ryerson University to lend this MRP to other institutions or individuals for the purpose of scholarly research. I further authorize Ryerson University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research. I understand that my MRP may be made electronically available to the public.

Acknowledgements

The majority of the database was provided by Dr. Richard R. Shaker, and remain his intellectual property for use and dissemination. Presence-absence data of aquatic invasive species and lake geometry was provided by the Adirondack Park Invasive Plant Program, a partnership program primarily run by volunteers and whose mission is to guard the Adirondack Park from invasive species impacts. With Dr. Shaker's hard work compiling a database, the volunteers from the plant program at Adirondack Park collecting data, studies such as these would not be possible and so my gratitude is extended to them. Additionally, my friends Mike and Krystal helped me with edits and presentation preparation so thank you so much for that, you are great friends.

Za mamu, tatu, sestru, i bake i deke.

Abstract

Aquatic invasive species, Eurasian Watermilfoil (EWM) and Curly-leaf Pondweed (CLP), have been dispersing across New York, USA and are threatening the ecosystem of Adirondack Park, a state park with a large forest preserve and heavily frequented by tourists. In this study, the prediction of EWM and CLP invasion across Adirondack Park lakes is modeled using logistic regression (LR) and geographically weighted logistic regression (GWLR) with lake, landscape, and climate variable predictors. EWM presence-absence is found to be best predicted by nearby invaded lakes, human presence, and elevation. The presence-absence of CLP models have similar findings, with the addition of game-fish abundance being important. GWLR increases model performance and prediction, with explained variation of EWM and CLP increasing by 23% and 16% and the percent correctly predicted increasing by 2.6% and 0.9%. The study shows that GWLR, a relatively novel methodology, works better than common LR models for predicting invasion of EWM and CLP across Adirondack Park, and corroborates anthropogenic influences on dispersal of aquatic invaders.

Table of Contents

List of Appendices

CHAPTER 1: Introduction

1.1 Introduction

Invasive species (IS) pose a serious threat to biodiversity, human health, and the economy worldwide, and their pressures are only growing with an increasingly globalized world (Lowe et al., 2000; Hulme, 2009; Butchart et al., 2010). The definition of IS by the U.S. government is "a species that is (a) non-native to the ecosystem under consideration; and (b) whose introduction causes or is likely to cause economic or environmental harm or harm to human health" (NISC, Executive Order 13112, 2017). The means by which a species becomes invasive is by passing geographical barriers to enter a new region, surviving in the new environment and creating self-sustaining populations. The species then must continue to spread and cause harm (NISC, 2006). Humans have been encouraging species spread by acting as transport vectors (Mack et al., 2000). In the United States (US), they have entered the country by accidental attachment to anthropogenic structures such as airplanes (e.g. Japanese beetle), automobiles and the ballast waters of boats (e.g. zebra and quagga mussels), and have even been deliberately introduced for ornamental/aquarium or agricultural purposes, though there have been some that naturally spread like the bolli weevil that came from Mexico (U.S. Congress, 1993; Miller, 2003). Once introduced to a region, they can continue to spread by similar means, such as fire ants who can cling to the mud flaps of trucks (U.S. Congress, 1993).

Aquatic invasive species (AIS) have and continue to be a large problem in the US, growing in number from 141 in 1900 to 1,161 in 2017 (USGS, 2017). They disrupt shipping activities, aquatic recreation, and clog up municipal water and industrial systems (Rockwell, 2003). Additionally, they can pose a health risk to humans by acting as vectors for viruses and enhancing disease epidemics. They can alter ecosystems by outcompeting other species and taking over other species' niches. There are potential benefits of AIS such as habitat for fish and invertebrates. Generally, however, the benefits do not outweigh the costs. Pimental et al. (2005) and the Office of Technology Assessment of the U.S. Congress (1993) estimate the total control and damage cost for AIS to be about $100 million USD annually. Note that ecosystem services (i.e. water purification, disease

regulation) were not taken into account in Pimental et al. (2005)'s damage estimate so it may be an underestimate of total damages (Lovell et al., 2006).

Identifying the means of spread across the US for AIS is of utmost importance if proper measures are to be taken to tackle their continued dispersion. In this study, these subsequent invasions from one invaded location to an uninvaded location after they have been introduced to a land mass will be called "secondary dispersal", as done so in Darbyson et al. (2009). The introduction of an invasive species does not necessarily imply they will prosper in the ecosystem. If the correct environmental conditions are in place for an AIS, only then will they be able to establish populations.

There have been a variety of methods employed to better understand the secondary dispersal of AIS. Some studies use observational and interview methods such as boat inspections and surveys of boat owners and bait shops (Johnson et al., 2001; Kilian et al., 2012). An understanding of the specific traits of invaded versus non-invaded lakes are another means by which to understand dispersal (Capers et al., 2007; Zanden and Olsen, 2008). Some more complex methods exist for predicting the pathways of spread such as gravity models, logistic regressions, and reaction-diffusion models (Hastings et al., 2005).

In the Adirondack Park Region of New York, there have been two pilot studies using logistic regression models to examine invasion risks to lakes (Shaker et al., 2013; Shaker and Rapp, 2013). Building from Shaker and Rapp's (2013) pilot study of the dispersal mechanisms of macrophytes Eurasian watermilfoil (EWM) and Curly-leaf pondweed (CLP) in 26 Adirondack lakes, and using an updated dataset from Shaker et al. (2017) for 126 lakes, this study seeks: (1) to explore lake, landscape, and climatic factors that may influence the spread and establishment of EWM and CLP; and (2) to explore a relatively new statistical methodology, geographically weighted logistic regression (GWLR), in its performance as a predictor model relative to a logistic regression. Knowledge of the combination of predictors that influence the presence-absence of these species would aid in understanding sites that may be vulnerable to invasion in the future (Zanden and Olden, 2008).

CHAPTER 2: Literature Review

2.1 Spatial Analysis for Environmental Management of Invasive Species

Spatial data processing was made commercially available in the 1980s with geographic information system (GIS) software, and environmental management agencies were one of the first to take advantage of this, recognizing that the environmental applications of GIS range from biodiversity evaluation to geological exploration (Goodchild, 2003). Spatial analysis is important for the management of IS spread and establishment as it is ultimately a spatial phenomenon. The simplest and most common question environmental managers will ask is: "where is the species found?" (Stohlgren and Schnase, 2006

Spatial analysis and modelling can provide an answer to this question as well as tackle where the species may soon be found as it spreads. Simple approaches to the spatial analysis of invasive species include mapping an invasion front, recording point locations of an IS to measure abundance across space (Hastings et al., 2005), or identifying dispersal mechanisms in the field (Johnson et al., 2001). Remote sensing can be also used to identify the occurrence and spatial structure of IS in an area (Walsh et al., 2008). Others have recognized that spatial heterogeneity of a landscape is important in understanding how a species will spread (With, 2002) and have incorporated it into their methods of estimating invasion speed and areas at risk to invasion (Prasad et al., 2010). Hotspot mapping of areas at risk to invasion are critical for IS monitoring programs (Hulme, 2009). For example, Drake and Lodge (2003) were able to find invasion introduction pathways from ship traffic patterns by using gravity models. Herborg et al. (2009) overlaid density maps of boating traffic with predicted maps of suitable habitats for an IS to create a hotspot map. There exist numerous more studies on predicting spread rates, risk of infestations, and estimating dispersal pathways (Higgins and Richardson, 1996; Buchan and Padilla, 1999; Neubert and Parker, 2004; Leung et al., 2006; Jimenez-Valverde et al., 2011). These models and analyses can help environmental managers understand the extent to which a landscape needs to be monitored or restored to tackle IS establishment and spread (With, 2002).

Methods of spatial analysis are also important for predicting invasion because they can provide information about the exact location that species may be found based on global

relationships or localized ones (applicable to only the study area). Zanden and Olden (2008) identified the questions that spatial analysis can help answer on the topic of secondary dispersal of macrophytes (Figure 1). If a certain set of predictors are found to be

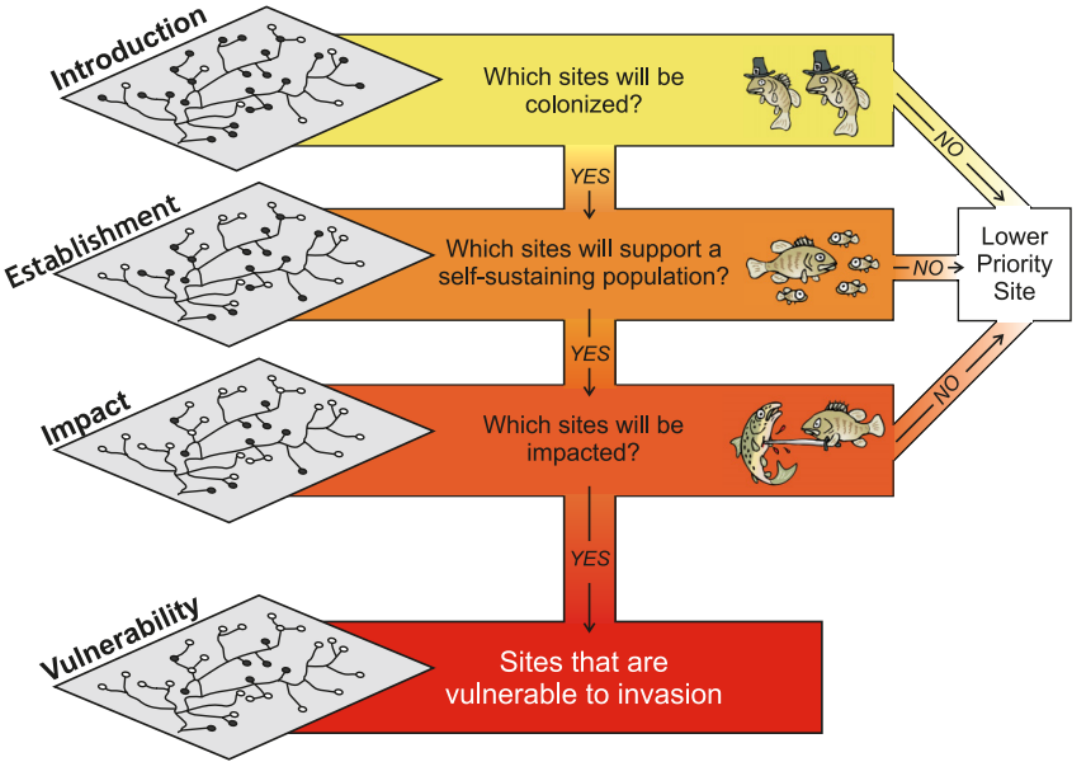


Figure 1. An assessment framework for the vulnerability of lakes to invasion. Figure by Zanden and Olden (2008).

important at predicting invasion in the study area, then it is known where invasion will occur based on these predictors. In this study, the prediction of possible invasion sites by AIS will be studied using the regression technique, GWLR. It is a relatively new methodology with only 27 published papers found in a Web of Science and Google Scholar query with the term "geographically weighted logistic regression" (see References, page 90). The methods applicability across various disciplines is evident with topics such as predicting the occurrence of fire ignition (Rodrigues et al., 2014), landslides (Feuillet et al., 2014; Zhang et al., 2016), landmine risk (Schultz et al., 2016), bank erosion (Atkinson et al., 2003), cloud cover (Wu and Zhang, 2013), and disease (Carrel et al., 2011). There is room to grow with the use of this method within many fields, though the most widely applied use of the method has been for assessments of natural disaster risk. For invasive species management, there have been no studies utilizing the method.

Furthermore, there are many steps that need to be taken before environmental managers can make decisions on how to best manage invasive species. There first needs to be incentives for tackling invasive species problems, that likely come through laws or regulations. In the United States the National Invasive Species Act (NISA) of 1996 recognizes that AIS are introduced by recreational and commercial boaters and states that there need to be measures put in place in the entire country to stop this from occurring. The Act allows funding for research on AIS spread and has a ballast water management program. Under this Act, the US has created a National Invasive Species Council and an Aquatic Nuisance Species Task Force which help implement the Act and make sure federal agencies and non-governmental organizations are acting accordingly (NISIC, 2017). New York (NY) is managed by the Northeast Aquatic Nuisance Species Panel, and the state itself has its own advisory committee and council whereby the committee provides information on spread prevention (by boats and trailers) to the council. The state also has a regulation that prohibits the storage, transport, import, sale, purchase, and introduction of invasive species (NY DEC, 2017). Other federal policies indirectly safeguard invasive species spread, such as the Clean Water Act (1977) which has created standards for water quality and wastewater. Water pollution changes the state of ecosystems and can enhance invasive species spread, and so this Act has a major influence on invasive species management. It is clear that policies are place for invasive species management in the U.S. and so to act on these, citizen science and open data can be of tremendous help. The Adirondack Park Invasive Plant Program (APIPP) is a program under the New York State Department of Environmental Conservation (NYS DEC) that is in partnership with more than 30 organizations and has hundreds of volunteers yearly to address the issue of Adirondack Park's IS (APIPP, 2016). Every year they monitor and regulate lakes across Adirondack Park, and their collection of spatial data makes understanding the risk of invasion and secondary dispersal mechanisms possible. This data is publicly accessible and can be used for spatial analyses studies that help prevent the species from spreading.

2.2 Background on Eurasian Watermilfoil and Curly-leaf Pondweed

Myriophyllum spicatum or EWM is a well-known and pervasive aquatic invader across North America with origins in Asia (Moody et al., 2016). The first documented case

of EWM was in the District of Columbia in 1942 and subsequent cases for EWM presence were reported from all across the US throughout the 1940s (Figure 2) and found in New York in the 1950s (Couch and Nelson, 1985). There is no consensus for how it was introduced but it may have been through the aquarium trade, fishermen using it for packing their bait, or through ballast disposal (Couch and Nelson, 1985; Les and Mehrhoff, 1999).

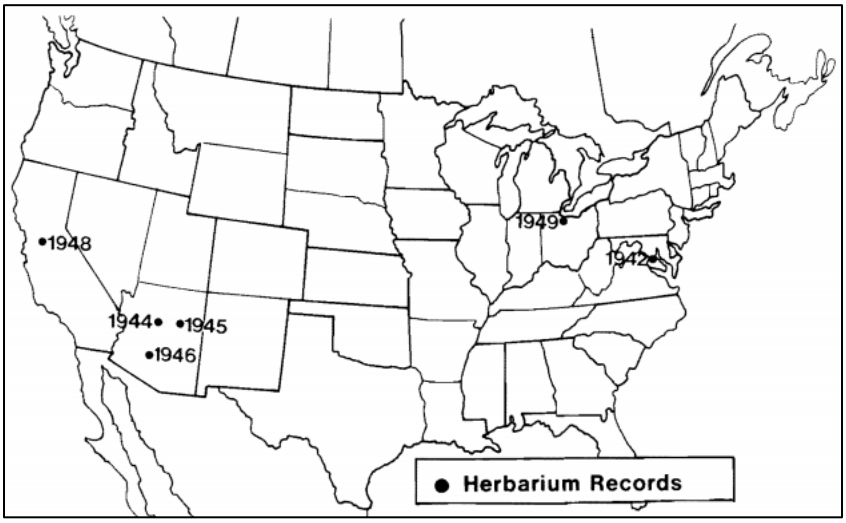


Figure 2. Earliest records of EWM in the U.S. prior to 1950. Figure from Couch and Nelson (1985).

Its successful invasion across New York, and generally the entire continent, has been due to its ability to thrive in a wide range of conditions, earlier spring blooming relative to other native aquatic plants, and ability to disperse by single-stem transport (Madsen et al., 1991). There are instances when EWM has been limited in its ability to invade when competing with certain other plants, such as water stargrass found in high abundances in northern Lake Cayuga, New York relative to EWM (Zhu and Georgian, 2014). Certain moths and beetles have also been found to disparage the growth of EWM (Johnson et al., 2000). Generally, however, EWM invasion has been successful in New York. For example, Boylen et al. (1999) examined the spread of EWM over 11 years from 1986 - 1997 in Lake George, New York and observed EWM eliminate 65% of the native species and those 35% remaining were severely reduced in abundance.

Potamageton crispus or CLP is another successful AIS across North America. The first reported incidence of CLP was in Wilmington, Delaware in 1859 (Stuckey, 1979). The plant spread to New York first in the Finger Lakes region during the mid-1880s and was later found in eastern New York's Lake George in 1897. The introduction of CLP may

have been through the aquarium trade for feeding waterfowl, aesthetics, or in fish hatchery stocks (Stuckey, 1979). It has owed it success to asexual reproduction, colonization techniques, and tolerance to varying aquatic conditions (Nichols and Shaw, 1986). See Figure 3 for a map of the earliest observations of CLP.

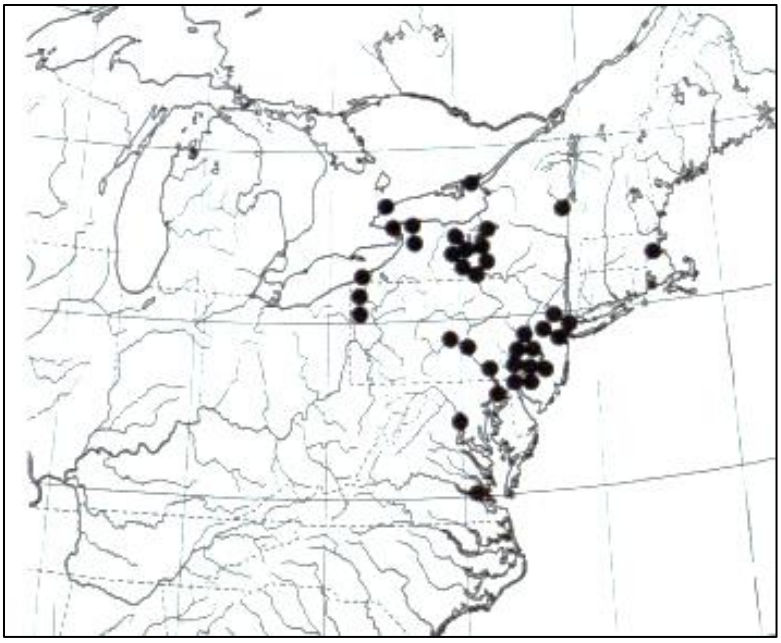


Figure 3. Earliest reported incidences of CLP prior to 1900. Figure from Stuckey (1979).

EWM has been a more aggressive aquatic invader than CLP as suggested by the data collected by Les and Mehrhoff (1999) which shows the number of new localities where the species has found in the US over time (Figure 4). The authors suggest that the most likely means of their spread was through escaping cultivation and that they continued to spread by vegetative propagules.

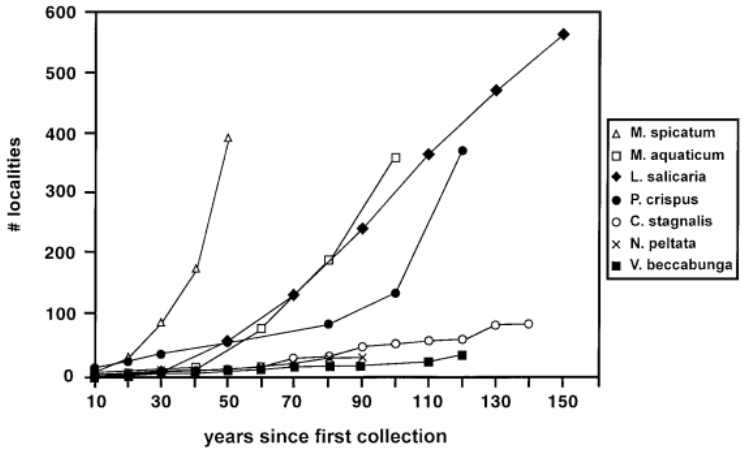


Figure 4. Les and Mehrhoff (1999) graphs showing the number of localities where AIS has been found since their first reported case.

There are various vectors for the secondary dispersal of these species, including natural means like wind and wave action that allow the plants to transport downstream or by waterfowl that eat CLP seeds (Catling and Dobson, 1985). A well-known vector for secondary invasions is by trailered boats (Rothlisberger et al., 2011). The term "boats" refers to a wide range of water vessels including power boats, fishing boats, sail boats, kayaks, canoes, and pontoons. AIS can cling to boat hulls and bilges, or be found on boat accessories like anchors, fishing gear, motors, live wells, recreational equipment and bait buckets (Johnson et al., 2001; Darbyson et al., 2009; Kilian et al., 2012; Bruckerhoff et al., 2015). They may also spread by being used for ornamental purposes in aquatic gardens and then escape cultivation or are dumped into waterways (Keller and Lodge, 2007).

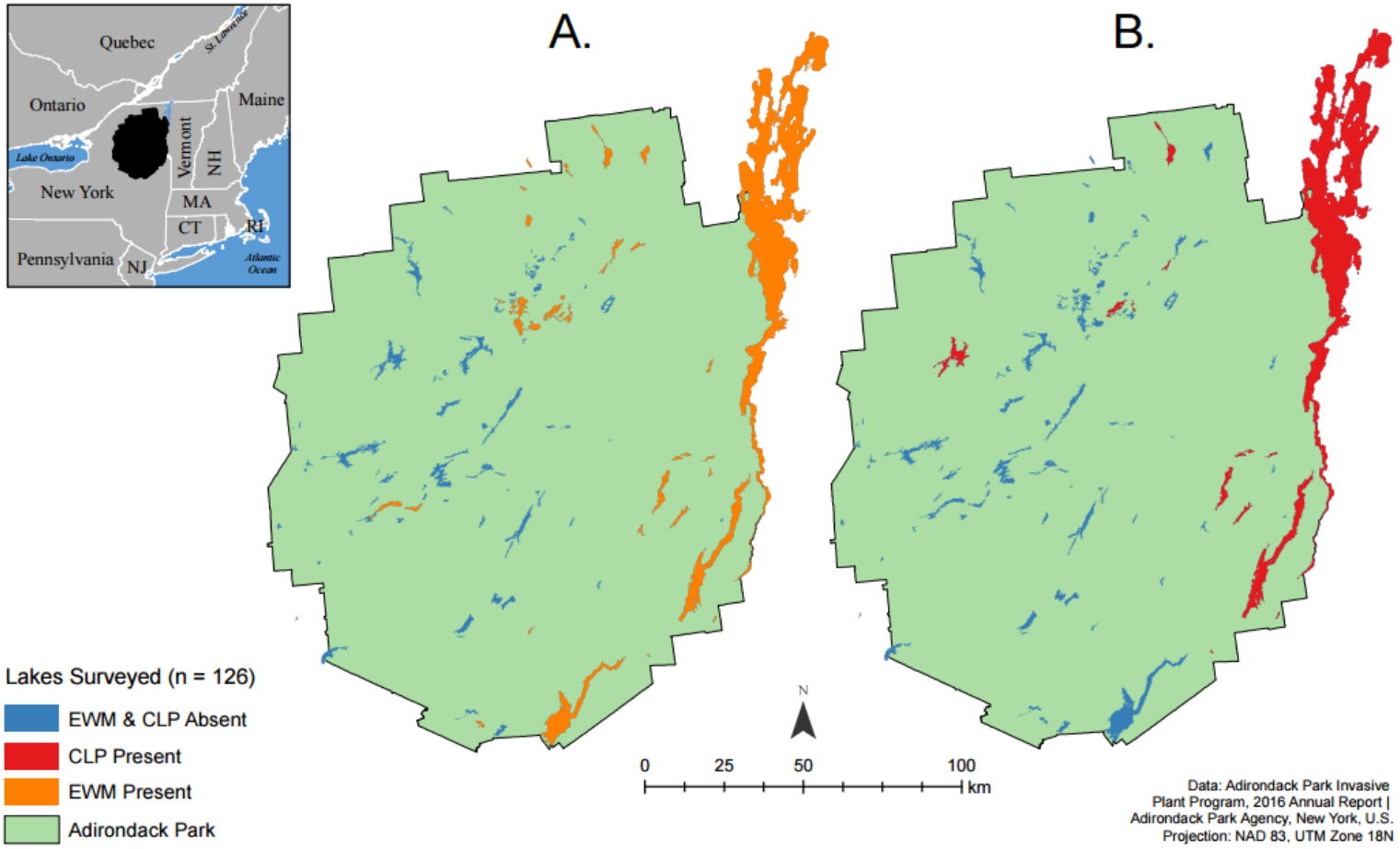
CHAPTER 3: Data and Methods

3.1 Study Area

Adirondack Park has over 3000 lakes, though only 126 lakes will be used in the study (APA, 2017; Figure 5). EWM is present in 43, or 34%, of the study lakes (Figure 5A) and CLP is present in 15, or 12%, of the study lakes.

The Adirondack Park encompasses 24,281 square kilometers and was created in 1892 to protect the forest against intense resource extraction occurring in the 1800s (Jenkins, 2004). Within the Adirondack Park is the Adirondack Forest Preserve, state lands belonging to New York (and thus the public) and making up 43% of the Adirondack Park (New York State, 2017). Forest preserves in New York are governed by Article XIV of the New York State Constitution which states that the lands be kept as wild forest, and that "they shall not be leased, sold or exchanged, or be taken by any corporation, public or private, nor shall the timber thereon be sold, removed or destroyed." Due to such strict measures, the forest has been left unaltered or undeveloped. The forests are made up of deciduous trees (3.53 million acres), evergreen forest (0.86 million acres), and mixed forest (0.60 million acres) (Adirondack Park Agency, 2017). The remaining land is privately owned for forestry, agricultural, or recreational purposes and is home to about 130,000 people (Jenkins, 2004; APA, 2017). The recreational tourism industry is quite large bringing in about $1.3 USD billion in 2015 (Oxford Economics, 2015).

Figure 5. Adirondack Park shown in black in inset map. Adirondack Park lakes (n = 126) show (a) presence-absence of eurasian watermilfoil and (b) presence-absence of curly-leaf pondweed.



Of the lakes in Adirondack Park, all bodies of water are public but the lands may be private and in fact seventy-five per cent of these lands are owned by the public but the remaining 25 percent are private (Jenkins, 2004). Lake Champlain is the largest of the lakes with a surface area of 1,331 square kilometers and is situated not only in New York but also Vermont in the US and the province of Quebec, Canada. The lake is a source of drinking water for about 200,000 people, is home to 81 species of fish, 318 birds, and 71 islands within it (LCLT, 2017). The second largest lake is Lake George found south of and connected to Lake Champlain by La Chute River. These lakes are part of the greater Lake Champlain watershed. Other basins found within Adirondack Park include the St. Lawrence, Oswegatchie/Black, Mohawk, and Greater Upper Hudson (APA, 2017).

3.2 Data

The presence or absence of EWM and CLP in Adirondack lakes will be used as the response variable and the data will be extracted from the APIPP Annual Report in 2016, which is inclusive of monitoring up until the end of 2016. The APIPP program has successfully conducted surveys on more than 300 lakes and ponds since the year of 2002. Of the monitored lakes, only those that have areas larger than 25 hectares are included for consistency in the dataset so this reduces the sample size to 126 lakes.

Furthermore, lake morphology traits are included in the dataset comprised of surface area, perimeter, perimeter-area ratio, maximum depth, and surface elevation. All variables are accepted for use in the model given their potential predictive abilities.

Overland transport of AIS can occur through boating activities so any variables that are representative of anthropogenic activity are good indicators of potential overland dispersal (Zanden and Olden, 2008; Johnson et al., 2001). This is reflected in the land cover data as access type, urbanized areas near lakes, distance to the interstate highway exit I-87 and nearest populated place.

Land cover percent for the riparian zone derived from the USGS National Land Cover Database 2011 is available in the dataset. These attributes are included in the model for this study because riparian zone conditions may be reflective of water quality in lakes, which in turn effect the habitat available for EWM and CLP. Forest and grass land cover in the zone can help to prevent or filter out nutrients from urban development or agricultural

areas from entering the lake. If the riparian zone is degraded, it may cause eutrophication conditions. Agriculture can cause excess nitrogen to enter a lake, and urban development can increase phosphorus flows (Carpenter and Cottingham, 1997).

Landscape metrics are also included, comprised of Shannon's diversity index (SHDI), Shannon's evenness index (SHEI), and relative patch richness (RPR). Diversity measures the amount of information for a species, patch type, or in this study, land cover, and is based upon richness and evenness. Richness is a measure of the number of patch types found in the landscape and RPR is measured as the percent of maximum potential richness given a user-specified number of patch types. Evenness measures the observed level of diversity of the total and SHEI is based on SHDI, which is a measure of land cover proportional abundance multiplied by the natural log of that value (McGarigal, 2015).

Additionally, class metrics are also part of the dataset. Class metrics include aggregation index (AI), percentage of like adjacency (PLADJ), area-weighted mean patch area (AREA_AM), and area-weighted mean Euclidean nearest neighbor distance (ENNMN) for developed open space and evergreen forest percent cover in the riparian zone. Area weighted-mean is a more accurate measure of the landscape than a simple mean as it takes into account the total percent covered by the patches and weights them accordingly when calculating the mean. It is reflective of the mean conditions of a pixel in the landscape, not a patch, when selected at random so it is called a "landscape-centric metric" (McGarigal, 2015). Euclidean nearest neighbor distance is an isolation metric that is measured as the distance to the nearest neighboring patch of the same land use type based on the shortest straight line (Lee et al., 2009; McGarigal and Marks, 1995). Both AI and PLADJ are landscape texture measures, or measurements of the propensity of patch types to be spatially aggregated, and dispersion measures, meaning they focus on how spread/dispersed a certain patch type is instead of how the landscape is intermixed (McGarigal, 2015). PLADJ is a spatial metric defined as the percentage of cell adjacencies for a certain patch type that are like adjacencies, cells bordering cells of the same patch type. It is equal to the summed number of like adjacencies for a land cover type by total number of cell adjacencies in the entre land cover. A higher value indicates more aggregation for that patch. Aggregation index (AI) is defined as the observed number of

like adjacencies relative to the maximum possible number of like adjacencies given the landscape composition (McGarigal, 2015).

These metrics are derived from the study of landscape ecology that evaluates how environmental patterns impact ecological processes. There are various ways to study landscape ecology, including composition/cover (aspatial) and configuration (spatial) approaches. They are complementary to one another as aspatial measures can provide the type and amount of land cover in an area, while a spatial measure will provide information about how the land is structured (Guerry and Hunter Jr., 2002).

Landscape and class metrics for two land cover types, developed open space (DO) and evergreen forest (EF), are in the dataset. Both are chosen to be included in the model because of their predictive abilities identified in past research. Previous studies of land use and their relation to ecological processes/conditions have used class and landscape metrics including entropy, Shannon's index, richness, mean patch size, dominance, contagion, fractal geometry, AI, and PLADJ (Guerry and Hunter, 2002; Alberti and Marzluff, 2004).

Climatic variables, temperature and precipitation, were added as predictors. The data was obtained from the U.S. Natural Resources Conservation Service as a raster dataset calculated from point data and a digital elevation model by parameter-elevation regressions on independent slopes model (PRISM) from 1981 – 2010 (USDA, 2017). Temperature influences lake ice cover, as temperature has been found to positively correlate with lake ice parameters (Williams et al., 2004). Precipitation also has an influence on lake characteristics, such as water clarity (Rose et al., 2016), and interacts with elevation as lakes higher receive more total water than those lower in a landscape (Kratz et al., 1997), so it may be a valuable predictor.

All data were acquired from Shaker et al. (2017) with the exception of climate variables. See Table 1 for the descriptive statistics of each variable and Appendix A for all variables data source.

Table 1. Mean, medium, standard deviation, and range (minimum – maximum) for non-transformed variables.

Variable	Units	Mean ± S.E.	Median	Standard Deviation	Min - Max
Response Variables					
Presence-absence of *Myriophyllum spicatum* (Eurasian watermilfoil)	0) absent 1) present	---	---	---	0 – 1
Presence-absence of *Potamageton cripus* (Curly-leaf pondweed)	0) absent 1) present	---	---	---	0 – 1
Predictor Variables					
Lake Morphology					
Lake area	sq. km	14.1 ± 9.0	1.5	100.5	0.2 - 1122.5
Perimeter	km	30.5 ± 8.6	12.6	96.9	1.9– 1037.4
Perimeter-area ratio	km/sq. km	8.1 ± 0.4	7.5	4.0	0.9 - 19.9
Maximum depth	m	18.1 ± 3.2	13.7	35.9	1.5 - 402.9
Surface elevation	m	454.0 ± 9.4	479.5	105.9	29.9 - 651.7
Other Lake Traits					
Access type	1) carry down only 2) public launch	---	---	---	1 – 2
Distance to invaded lake					
Eurasian watermilfoil	km	14.9 ± 1.0	12.2	11.4	1.3 - 45.1
Curly-leaf pondweed	km	28.9 ± 1.7	25.7	18.6	1.3 - 94.9
Game fish abundance: yellow perch, smallmouth bass, rainbow trout	0) absent 1) one species 2) two species 3) three species	---	---	---	0 – 3
Distance to I-87 exit	km	59.3 ± 2.4	59.7	26.6	2.6 - 116.8
Distance to nearest populated place	km	3.5 ± 0.2	3.1	2.2	0.2 - 13.6
Climate					
Average temperature	°C	5.5 ± 0.1	5.3	0.7	4.4- 8.2
Maximum temperature	°C	11.7 ± 0.1	11.5	0.6	10.4 - 14.7
Minimum temperature	°C	-0.9 ± 0.1	-1.1	0.8	-2.2 - 2.1
Range temperature	°C	12.3 ± 0.04	12.4	0.474	10.9 - 13.5
Average precipitation	inches	44.3 ± 0.3	44	2.9	37.8 - 53.5
Maximum precipitation	inches	45.1 ± 0.3	44.3	2.9	39.8 - 55.6

Minimum precipitation	inches	44.5 ± 0.3	44.1	3.03	33.4 - 53.2
Range precipitation	inches	0.6 ± 0.1	0.2	1.3	0 - 10.4
Land Cover ϕ	Percent of total area				
Developed open space	%	3.4 ± 0.4	2.1	4.4	0 - 26
Developed low intensity	%	0.5 ± 0.1	0.1	1.0	0 – 6.9
Developed medium intensity	%	0.2 ± 0.05	0	0.6	0 - 4.4
Developed high intensity	%	0.03 ± 0.01	0	0.1	0 - 0.8
Deciduous forest	%	22.4 ± 1.2	21.2	13.1	0 - 60.6
Evergreen forest	%	20.1 ± 1.1	19.9	12.7	1.2 - 60.6
Mixed forest	%	5.3 ± 0.5	3.9	5.2	0 - 32.9
Pasture and hay	%	0.1 ± 0.04	0	0.4	0 - 3.5
Cultivated crops	%	0.04 ± 0.01	0	0.2	0 - 1.08
Shrub and scrubland	%	2.2 ± 0.3	1.02	3.2	0 - 22.7
Herbaceous	%	0.2 ± 0.05	0	0.5	0 - 3.2
Emergent herbaceous wetland	%	0.8 ± 0.2	0.2	1.9	0 - 14.3
Woody wetland	%	3.5 ± 0.4	2.13	4.1	0 - 25.3
Open water	%	41 ± 1.07	39.4	12	16.4 - 83.8
Barren	%	0.4 ± 0.2	0	2	0 - 16.3
Land cover class metric ϕ					
AI, DO	%	47.5 ± 2.1	55.5	23.9	0 - 87
AI, EF	%	76 ± 0.6	76.6	6.6	54.1 - 90.9
PLADJ, DO	%	42.7 ± 2	49.5	22.4	0 - 81
PLADJ, EF	%	73.2 ± 0.7	74.3	7.3	48.5 - 88.5
AREA_AM, DO	sq. m	5.5 ± 0.7	2.5	7.9	0 - 55.8
AREA_AM, EF	sq. m	25 ± 2.3	14.6	25.6	1 - 135
ENN_AM, DO	m	144.8 ± 43.9	82.7	492.6	0 - 5480.8
ENN_AM, EF	m	111.1 ± 5.7	86.2	64.2	13.5 - 456.1
Landscape diversity ϕ					
RPR	%	57.04 ± 1.3	56.3	14.8	25 - 93.8
SHDI	SHDI ≥ 0, w/o limit	1.4 ± 0.02	1.4	0.2	0.8 - 2
SHEI	0 ≤ SHEI ≤ 1	0.7 ± 0.01	0.7	0.1	0.3 - 0.9

ϕ Calculated within a 300 meter buffer of the lakes.

3.3 Methodology

Logistic regression (LR) and geographically weighted logistic regression (GWLR) will be used to assess the impacts of several variables on the presence-absence of EWM and CLP. LR is a generalized linear model (GLM) and global approach while GWLR is a local approach and extension of LR that accounts for non-stationarity in the model. These two models will be compared in their performance as predictive models.

3.3.1 Logistic Regression

Linear regression is similar to LR, as it is a GLM, and will be briefly explained. Linear regression aims to determine which variables better fit the shape of the data relative to the mean by using a best fit line. The equation to fit a line to data is

$$Y_i = ax_i + b + \varepsilon_i \qquad \varepsilon_i \sim N(0, o^2)$$

where a is the slope of the straight line which has been fitted to the data, b is the intercept of the line, o is a fixed value for all data points and ε_i random error. This function allows for the prediction of y for any value of x (Orloff and Bloom, 2014). The regression coefficients are obtained by using a least squares methodology that finds the minimal summed distance between a line through data and the respective data points. A significant linear regression model is one that finds a smaller summed distance to the best fit line relative to the summed distance to the mean of the data (Field, 2009). This methodology can be extracted from a univariate relationship to a multivariate one with multiple predictors. The formula then takes the form

$$Y_i = b + a_1x_{i1} + a_2x_{i2} + a_3x_{i3} + \cdots a_rx_{ir} + \varepsilon_i \qquad \varepsilon_i \sim N(0, o^2)$$

where $x_1, x_2, \ldots x_r$ is a set of r predictors in the *i*th sample, a_i (i = 0, 1…. r) are regression coefficients. In linear regression the response and predictor variables are continuous, however, the function can be extended to apply to categorical response variables in LR (O'Reilly et al., 2007).

LR predicts the probability, not the value, of the response variable given a set of continuous or categorical predictor variables (Field, 2009). The LR probability of a response variable P(Y) is derived as follows

$$probability, P = \frac{outcome\ of\ interest}{all\ possible\ outcomes}$$

whereby an odds ratio is then derived from probability as

$$odds\ ratio = \frac{P(event\ occuring)}{P(event\ not\ occurring)}$$

$$odds\ ratio = \frac{P}{(1-P)}$$

To tie together the linear combination of predictor variables with a probability distribution, the natural logarithm of the odds ratio is taken, called the logit.

$$\ln(odds\ ratio) = ln\frac{P}{1-P} = logit(P)$$

The inverse of the logit function is taken to derive probabilities of 0 and 1 along the y axis.

$$logit(P) = \ln\frac{P}{1-P} = b + a_1x_{i1} + a_2x_{i2} + a_3x_{i3} + \cdots a_rx_{ir}$$

$$\frac{P}{1-P} = e^{(b+a_1x_{i1}+a_2x_{i2}+a_3x_{i3}+\cdots a_rx_{ir})}$$

The probability is then isolated in the equation, resulting in the estimated logistic regression equation,

$$P(Y) = \frac{e^{(b+a_1x_{i1}+a_2x_{i2}+a_3x_{i3}+\cdots a_rx_{ir})}}{1+e^{(b+a_1x_{i1}+a_2x_{i2}+a_3x_{i3}+\cdots a_rx_{ir})}}$$

where P(Y) is the probability of the response variable, Y, occurring, e is the base natural logarithm, and the bracketed portion is the same equation as a linear regression equation (Rupert et al., 2008). By using a logarithmic transformation of the linear regression equation, the assumption of linearity in the model is overcome (Field, 2009).

LR chooses regression coefficients, a_i (i = 0, 1…. r), based on a maximum likelihood procedure, unlike linear regression which uses a least squares procedure. All possible sets of regression coefficients are evaluated to see which one best fits the observed data and the one that approximates it closest is chosen. The evaluation is performed by a likelihood function, calculated as

$$L(b, a_i) = \prod_{i=1}^{n} p\,(x_i)^y [1 - p(x_i)]^{1-y_i}$$

then simplified by taking the logarithm,

$$ln\,[L(b, a_i)] = \sum_{i=1}^{n} y_i \ln[\,p(x_i)] + (1 - y_i)\ln[1 - p(x_i)]$$

where one can solve for the value of *a* to maximize *L (b, a)* by differentiation *L (b, a)* with respect to *b* and *a* and having the expression set equal to zero (Hosmer and Lemeshow, 2000; Shalizi, 2017).

3.3.2 Geographically Weighted Logistic Regression

GWLR works by using a moving window across the study region and performing localized regressions. Regression coefficients are spatially weighted and the size of the window is determined by a bandwidth specification appropriate for the study (O'Sullivan and Unwin, 2010). The GWLR work will be performed in GWR4 software which is the only stand-alone software for GWR modelling, though there are GWR packages in R such as *GWmodel* that can also perform this technique (Maynooth University, 2017).

A GWR model is explained by the equation:

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i$$

where i represents the index of each sample point, (u_i, v_i) are the spatial coordinates of each ith point, β is the beta or slope (β_0 is representative of the intercept), x is the covariate or independent variable, y is the dependent variable, k is the index of the covariates, and ε_i is an error term. In matrix form, GWR is

$$Y = (\beta \otimes X)1 + \varepsilon$$

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0(u_1, v_1)x_{11} & \cdots & \beta_k(u_1, v_1)x_{1k} \\ \vdots & \ddots & \vdots \\ \beta_0(u_n, v_n)x_{n1} & \cdots & \beta_k(u_n, v_n)x_{nk} \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

where β and X have the same dimension's $n \times (k+1)$, and Y will be a $n \times 1$ vector with n observation/data points and k independent variables (Fotheringham et al., 2002).

To estimate beta, $\hat{\beta}$, in the GWR model, the following is computed in matrix format:

$$\hat{\beta}(i) = (X^T W(i) X)^{-1} X^T W(i) Y$$

where i is a row of the matrix from the beta in the GWR equation, and W(i) is a spatial weight matrix with the diagonal elements holding the weight values and the non-diagonal elements being zero. There are numerous weighting procedures that can be used to find W(i), and those discussed here are taken from the options available in GWR4 software (Nakaya, 2016). The moving 'window' that is used in GWR is specified by the kernel type and bandwidth specification. In Figure 6 it can be seen how the window/kernel operates, where a regression point, i, is shown relative to observation points, j (Fotheringham et al., 2002).

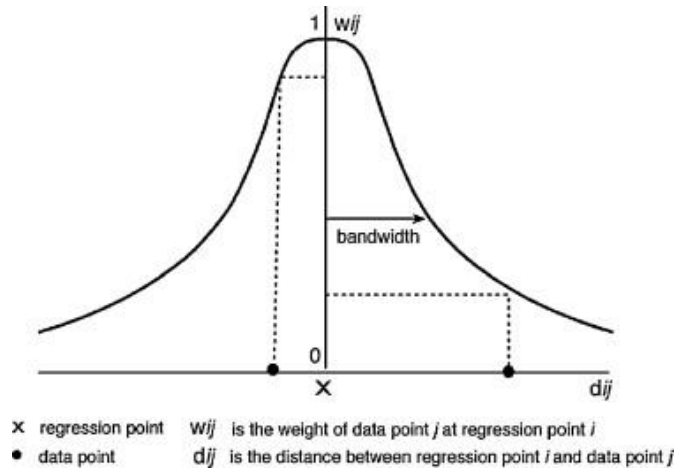


Figure 6. Weighting scheme in GWR as illustrated by Fotheringham et al. (2002).

The kernel can be fixed so that the kernel size (same number of data points or same distance) is the same for each local regression, or it can be adaptive and the size fluctuates to always include a certain number of data points (Charlton and Fotheringham, n.d.). For the kernel functions, bi-square or Gaussian options can be used. A Gaussian function weights data points smoothly by giving the highest weight to the middle of the kernel (at the regression point) and gradually decreasing weight away from the middle. A bi-square weighting function also weights data points by giving the highest weight at the middle and decreasing gradually however it specifies a distance at which weighting becomes zero (Nakaya, 2016).

Bandwidth selection is another important element in the weighting function to consider. It is a measure of distance decay within the weight formula. With a larger bandwidth (greater distance or number of data points), the results would be more smoothed and with a smaller bandwidth the results would be rougher (Fotheringham et al., 2002). There are optimal values to choose for a bandwidth which can be found by using different search algorithms including golden section search and interval search, or if the user is knowledgeable about the predictors and study area then they could specify the bandwidth themselves. The search algorithms increase the efficiency of finding the best model. See Figure 7 below for the weighting options available in GWR4 software and how they are calculated.

Fixed Gaussian	$w_{ij} = \exp(-d_{ij}^2 / \theta^2)$
Fixed bi-square	$w_{ij} = \begin{cases} (1 - d_{ij}^2 / \theta^2)^2 & d_{ij} < \theta \\ 0 & d_{ij} > \theta \end{cases}$
Adaptive bi-square	$w_{ij} = \begin{cases} (1 - d_{ij}^2 / \theta_{i(k)})^2 & d_{ij} < \theta_{i(k)} \\ 0 & d_{ij} > \theta_{i(k)} \end{cases}$
Adaptive Gaussian	$w_{ij} = \exp(-d_{ij}^2 / \theta_{i(k)}^2)$

Notes: i is the regression point index; j is the locational index;
w_{ij} is the weight value of observation at location j for estimating the coefficient at location i.
d_{ij} is the Euclidean distance between i and j;
θ is a fixed bandwidth size defined by a distance metric measure.
$\theta_{i(k)}$ is an adaptive bandwidth size defined as the k th nearest neighbour distance.

Figure 7. Weighting options in GWR4 software (from Nakaya, 2016).

During a bandwidth search process, the bandwidth is chosen by using a statistical measure of the model which will choose the "best" model. Statistical measures for this include Cross-Validation or information theory (IT) metrics Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC). Unlike null hypothesis significance testing approaches, IT is particularly useful when working with many predictors because it allows for one to have multiple competing models and compare them by the metric (Symonds & Moussalli, 2011). IT metrics are meant for ranking models and comparing them, and does not stand alone as a measure. When using AIC and BIC, models with the lowest value are taken to be the best model, however, it should be kept in mind that AIC and BIC metrics do not evaluate fit of the data (Burnham and Anderson, 2002). The criterion works on the principle of parsimony where it best attempts to balance bias and variance or under fitting and overfitting of the model taking into account the likelihood, L, and the number of fitted parameters, k. Those models with a higher k will have a higher IT metric value, so the criterion favours models with less parameters (Symonds & Moussalli, 2011). BIC works

similarly, but the formula is tweaked by multiplying k with the natural logarithm of n, which in turn has the impact of selecting models with even fewer parameters. Cross validation methods work by picking the model with the best predictive capabilities and a more detailed description of it can be found in Fotheringham et al. (2002). Simpler goodness of fit measures would not be appropriate for model/bandwidth selection because they would always prefer models with a very high number of parameters which may cause overfitting.

In this study, the centroid of lakes as specified by ArcMap software was used in the GWLR weighting procedure and specifications selected as deemed appropriate. Since the lake centroids are irregularly spaced data points, a fixed kernel type would not be appropriate to use because some regressions would include many more data points within them compared to other regressions in the study region and so an adaptive kernel type was used to have consistent regressions. For the kernel function weight, a Gaussian weighting option was used because many predictors may have an influence over a wide area. EWM and CLP may get transported by overland dispersal accidentally on boats, trailers or fishing equipment and so dispersal events may be not localized but may "jump" around depending on where humans travel. To specify a range at which weighting ceases is likely inappropriate due to these potential mechanisms, and so a Gaussian kernel weight was used. The bandwidth will be determined by using a golden section search since the study is exploratory, and the study area not well known. A golden section search works by finding extremums of criteria selection valuess and searching between them for the lowest number (see Fotheringham et al., 2002 for a more thorough understanding).

AIC will be the statistic used for finding an optimal model because it is a commonly accepted and used measure, and was used in Shaker et al. (2017). Additionally, a maximum of six predictors will be accepted for any model so BIC may not benefit model selection as six is a small enough number of predictors. Overfitting can be avoided by having a smaller number of predictors as LR has been found to need at least 10-15 observations per predictor (Babyak, 2004). AIC is derived as

$$AIC = -2\ln(L) + 2k$$

although for small sample sizes (when n/k < 40 with n/k being derived from the most complex model and k as the number of fitted parameters), the corrected version (AICc) is more appropriate and calculated as

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1}$$

For AIC, a difference of +/- 3 values of AIC between models is considered substantial enough to show there is a difference between the models.

The results of GWLR include local (for each lake) beta values, t-statistic, and R_L^2 which provide insight on regions with high statistical associations or areas without any. The global LR beta values may be significant, but they do not provide associative information (AIS presence-absence with the predictor) about specific areas within the study region. The best way to see local associations is by mapping the beta values. On the maps, the t-value is used to show lakes that are not statistically significant at the 90% level; lakes with t-value between ± 1.96 have no statistical association (Matthews and Yang, 2012).

There are potential limitations to the GWLR methodology that should be considered. Since it employs a localized approach, the number of observations included in each local model may be small which can lead to greater error, and the potential to overestimate variation in space that may not actually exist. However, the model offers a novel approach in taking account spatial variation without using spatial coordinates as covariates in a simpler regression, which may confound explained variability with variables like precipitation or temperature (Zuur et al., 2009).

3.3.3 Data Reduction and Meeting Regression Assumptions

Prior to performing LR and GWLR, there are model assumptions that must be assessed. Firstly, given that logistic models are an extension of linear regression, the predictor is assumed to have a linear relationship with the logit (Field, 2009; Stoltzfus, 2011). This can be tested once a model is built but to better ensure this is the case, an evaluation of normality was undertaken for continuous variables by examination of frequency distribution histograms, Q-Q plots, descriptive statistics (central tendency, variability measures), and measurements of shape (skewness, kurtosis). Skewness

measures the symmetry about frequency distributions of a dataset, where perfectly symmetrical distributions have a skew of zero (Field, 2009). Kurtosis is a measurement of the degree to which the clustering of data differs to a normal frequency distribution (Hopkins and Weeks, 1990; Field, 2009). Also, a Shapiro-Wilks test was performed for all transformed variables to assess normality, as it has been found to be the most powerful relative to other tests (Razali and Wah, 2011). Those data exhibiting non-normality were transformed by commonly used means, including logging and square rooting (see Appendix B; Erickson and Nosanchuk, 1992; Cheruvelil and Soranno, 2008; Shaker, 2008). Outliers were not discarded as they may be indicative of spatial phenomena that may help to explain presence-absence of AIS in a particular lake. Shapiro-wilks test found all variables to be statistically significant ($p<0.05$) except for AREA_AM, EF and SHEI, so there is evidence that the variables are not normally distributed. However, given that outliers were not discarded which may have significantly skewed or kurtosed the data, and that linear regression models are quite robust against small deviations from normality (Zuur et al., 2009), the predictors are still accepted for use in the global and local models. Those continuous variables that had extreme deviations from the normal distribution (generally skew values greater than ± 1 and kurtosis greater than ± 3), and/or those that would contribute little to the model (i.e. land cover ≤ 1%), were taken out of the model.

The second assumption is that there is no multicollinearity or correlation between predictors present (Field, 2009). Collinearity is an extremely important assumption to meet since the study's primary goal is to explore which covariates contribute to predicting the response variable and this is only possible if the variables are independent of one another (Zuur et al., 2009). Correlation coefficients and variance inflation factors (VIFs) were computed in R statistical software. Pearson Product-moment coefficients (R_p) were evaluated for a subset of continuous predictor variables. For all the variables, a Spearman's correlation, R_s, was used because this statistic accommodates both continuous and ordinal type data (Shaker et al., 2017). A threshold of $|R_s| \geq 0.75$ was used (Shaker, 2013; Shaker et al., 2017), though in other studies a threshold of 0.7 is used (Fielding and Haworth, 1995; Dormann et al., 2013) so both are considered. Pearson correlations amongst lake geometry and climatic variables revealed strong correlations as expected by value and scatterplot analysis (Appendix C.1). As such, area, maximum temperature and precipitation, and

minimum temperature and precipitation were taken out of future models. Area was chosen to be removed from the model instead of perimeter because it had higher overall correlation coefficients with all other lake morphology variables in comparison to perimeter.

The Spearman's rank correlation coefficient identified numerous correlations greater than 0.7 and 0.75 (Appendix C.2). Open water and perimeter-to-area (P-A) ratio had an R_s = -0.82 so it was decided that the P-A ratio was to be dropped because open water within the 300 meter buffer of the lake may be an important predictor for the spread of AIS as freshwater connectivity is a natural means of spread for any macrophyte. Developed open space and AI, DO had high correlations with AREA_AM, DO so AREA_AM, DO was dropped from the model. Developed, low intensity had a correlation with developed, medium intensity of 0.743. It was also correlated with RPR (R_s = 0.729) so it was dropped from the model. Percent open water and SHEI have an R_s = -0.732, however neither were chosen to be dropped from the model because they are still below the 0.75 threshold, but more importantly because they are likely to be important predictors. Both PLADJ, DO and PLADJ, EF measures had the highest R_s values (PLADJ, DO and AI, DO R_s = 0.953; PLADJ, EF and AI, EF R_s = 0.983) so were removed from the model. Percent evergreen forest and its associated landscape metrics were all correlated above 0.7, so it was decided to drop percent evergreen Forest and Area_AM, EF and keep ENNAM, EF and AI, EF as representative of evergreen forest in the landscape.

VIFs were calculated as

$$VIF = \frac{1}{1 - r_i^2}$$

where r_i^2 is the coefficient of determination defined as

$$r^2 = \frac{sum\ of\ squares\ due\ to\ regression}{total\ sum\ of\ squares}$$

The VIF equation computes high values when the coefficient of determination is high, so when r^2 = 0.9, VIF = 10. Therefore, for VIF values the maximum threshold is 10 (Dormann

et al., 2012), VIF > 5 is a signal of concern (Shaker, 2015), and preferred are VIF < 2.5. Stepwise VIF regression analysis was undertaken by a pre-written function in R statistical software (see Appendix C.4 for the code). Variables with VIF > 10 were removed including PLADJ_EF, PLADJ_DO, SHDI, percent evergreen forest, and AREA_AM EF. Twenty-six variables remained in both the EWM and CLP data after the multicollinearity analyses.

Another form of collinearity is spatial autocorrelation, the phenomenon that data closer in location are more similar to each other than those further apart (Tobler, 1970). Spatial autocorrelation is important to test for in the data because it accounts for any significant spatial patterns in the data, and is a common environmental phenomena. Global Moran's *I* was used to test for spatial autocorrelation and was calculated in ArcMap 10.4.1 software. The Moran's *I* statistic looks at whether the data is clustered, dispersed, or randomly distributed in space by inference of a null hypothesis (ESRI, 2017). It is calculated as

$$I = \left[\frac{n}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right] \times \left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}\right]$$

where wij is the spatial weights matrix, n the number of observations, yhat the mean, and yi the observation in one zone and yj the observation in the neighboring zone. The "neighbor" is determined by the user depending on what best suits the data and the user's needs. Neighbors can be chosen based on contiguity, distance, or triangulation methods (O'Sullivan and Unwin, 2010). A positive Moran's *I* indicates that neighbors have values on the same side of mean, while a negative one means that the neighbors have opposing values relative to the mean. If the statistic has a significant p-value then the data may be clustered or dispersed in space (ESRI, 2017). Similarly, a local Moran's I was calculated for presence-absence response variables with positive values indicating clustering and negative ones, an outlier. Furthermore, a join count statistical test was used to test for spatial autocorrelation of all categorical variables. This was done for presence-absence, access type, and game fish data. The test counts the occurrences of neighbor pairs and compares it to an expected count by analytical or permutation inferences (O'Sullivan and Unwin, 2010). The join count test was run in R software using the package spdep, with the

functions joincount.mc for the permutation and joincount.test for the inferential test. A BB join was measured which is the number of similar neighbor joins that are observed and compared to an expected amount by a z test statistic (observed – expected / standard deviation of expected joins).

3.3.4 Model Building

Once these assumptions were addressed, models needed to be built for inclusion in the LR and GWLR. The prior multicollinearity analysis already played a role in reducing the number of variables to include by discarding those that caused collinearity issues. Henceforth, bivariate logistic regression for each predictor was performed to identify which variables showed statistically significant relationships with the presence-absence of the AIS of interest. To evaluate the contribution of each predictor variable, the Student's t and Chi-square was computed and evaluated for significance. Student's t is a measure of whether the data yielded a coefficient due to sampling error or not (NCES, 2017). The Chi-Square statistic was also evaluated, which is a statistic measuring the difference between the model with the predictor variable versus the model with only the constant. This was calculated in SPSS as

$$Chi - Square\ Test\ Statistic = -2 \ln \left[\frac{likelihood\ without\ the\ variable}{likelihood\ with\ the\ variable} \right]$$

which can also be called the deviance statistic or likelihood ratio test but it follows a chi-square distribution so it takes on that name in SPSS (Hosmer and Lemeshow, 2000). Those variables that did not show significance of at least 90% for both diagnostic tests were dropped from inclusion in the model.

Those variables that were significant in the bivariate logistic regression models were used to build three to five different possible models for the prediction of EWM and CLP. As mentioned previously, a limit of six predictors per model was accepted to avoid overfitting but additionally to simplify interpretation. The "best" model was chosen by AIC. IT-AIC selection method was employed in two different ways. Firstly, the significant variables from the bivariate logistic regression were placed into an automated model

selection algorithm to find the best global logistic regression models through an exhaustive screening approach that picked the model with the lowest AICc. This was performed in R statistical software with the *glmulti* package (Calcagno, 2015; Calcagno and Mazancourt, 2010). Secondly, a manual stepwise AICc procedure was undertaken in GWR4 statistical software with the same significant variables from the bivariate logistic regression for an exploration of other potential models (Fotheringham et al., n.d). The final selection of models from these two processes made sure to include the global model with the lowest AICc and to choose alternative models with non-redundant variables (as deemed appropriate based on predictor contributions and background research).

Once a subset of models was chosen, the residuals for each model were examined. Residual diagnostics are important to evaluate for LR however, the assumptions that are present in linear regression do not hold for LR. In linear regression, it is assumed that the errors are normally distributed with a zero mean and constant variance. Since the response variable is binary in LR, the residual will follow a binomial distribution with a zero mean and variance of P(Y)*(1-P(Y)) (Hosmer and Lemeshow, 2000). Thus, a test of normality was not carried out for the residuals. On the other hand, it is still important to examine the residuals for outliers, those points that do not fit the model well. This is because residuals in a GLM are expected to be independent of one another and correlation can violate this assumption and cause the model to under- or over-estimate the predicted probabilities (Fotheringham et al., 2002). Pearson residuals were calculated, as

$$Pearson's\ residual = ZRE = \frac{Y_i - \hat{Y}_i}{SQRT\left(\hat{Y}_i\left[1 - \hat{Y}_i\right]\right)}$$

where Y_i is the observed outcome and $\hat{Y}_i$ is the predicted probability. In conjunction with this, Global Moran's I calculation is important for showing if any spatial patterns exist that may indicate spatial autocorrelation (Gumpertz et al., 1999; Fotheringham et al., 2002).

3.3.5 Model Diagnostics

A primary goal of this study is to evaluate the performance of GWR and compare it to a GLM. To meet this goal, the two modelling methodologies must be compared by

their goodness-of-fit and predictive capabilities. Previous papers on the topic have utilized likelihood measures, AIC, Bayesian information criterion, percent deviance, area under the curve, percentage correctly predicted, spatial autocorrelation of residuals, residual errors, and pseudo R^2 (Carrel et al., 2011; Feuillet et al., 2014; Luo and Wei, 2009; Martinez-Fernandez et al., 2013; Wimberly et al., 2008; Windle et al., 2010; Zhang et al., 2016).

To evaluate the goodness-of-fit of the models, common diagnostics will be used. A pseudo R_L^2 measure, McFadden's rho-squared/log likelihood ratio, was calculated for evaluating model adequacy (McFadden, 1973). It is called a 'pseudo' R^2 because it is analogous to R^2 from a linear regression model since its values also range from 0 to 1 and are interpreted in an akin way to R^2. It is calculated as

$$R_L^2 = -\frac{[\ln(L_0) - \ln(L_M)]}{[\ln(L_0)]}$$

where L0 is the likelihood function for the model with just the intercept, Lm the likelihood function for the model with all predictors (Menard, 2000). Another measure that will be reported is a likelihood measure (or deviance), -2 log-likelihood (-2LL), which was previously explained. A perfect fit to the data would be a value of zero for -2LL so the lower the value, the better the fit. This measure is analogous to the total sum of squares in a linear regression (Hair Jr. et al., 2010). Lastly, the AICc values that were used for model building will be examined as a means of comparing model performance.

Furthermore, a measure of predictive accuracy will be included. The percentage of correctly classified cases will be evaluated with a cut-off of 0.5. For $P(Y) > 0.5$ with $Y = 1$ and $P(Y) < 0.5$ with $Y = 0$, they are counted as correctly classified and those not meeting these conditions are not correct classifications. This is a simple measure that will only be used if the majority of P(Y) do not fall around 0.5 which would provide misleading results. Should this be the case, the receiver operating characteristic (ROC) curve will be used instead as suggested by Hosmer and Lemeshow (2000) and reported as a commonly used means of evaluating predictive ability (Peng and So, 2002; Stoltzfus, 2011). Another measure of predictive power will be assessed from the residuals, the root mean square error

(RMSE). The RMSE is the mean of the absolute residuals, thus the further the value is from zero, the worse the model is and vice versa at prediction (Feuillet et al., 2014).

Lastly, a key part of assessing the performance of a GWR model against a GLM is its ability to diminish spatial autocorrelation (Fotheringham et al., 2002). The Global Moran's I Pearson residuals from the LR will be compared to those of GWR.

The methodological steps are summarized below in Figure 8 as the process is complex and a diagram will help to showcase the steps more simply.

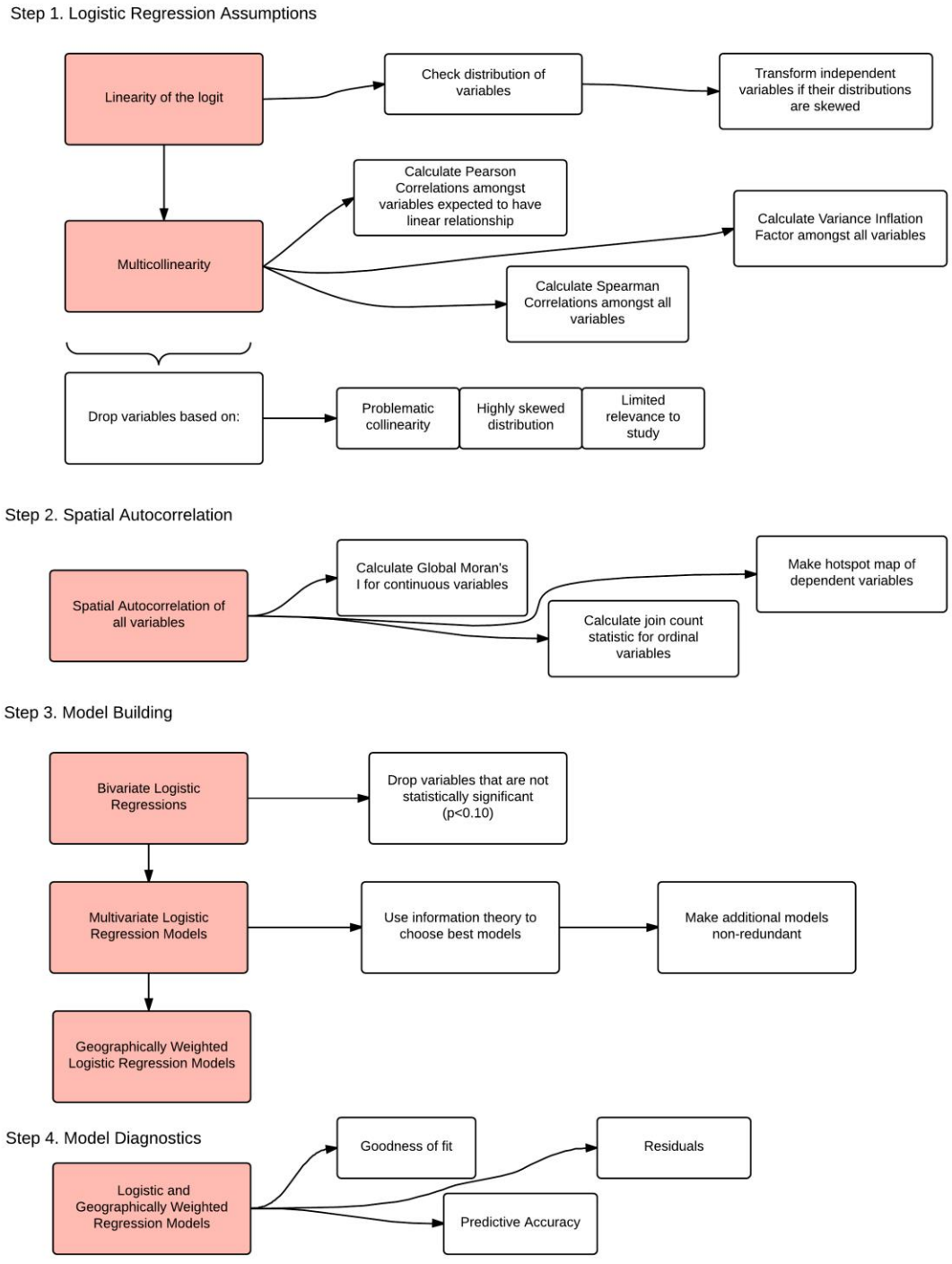


Figure 8. Methodology of study. Pink = name of procedure. White = subheadings for carrying out the procedure.

CHAPTER 4: Results

4.1 Spatial Autocorrelation

The Global Moran's *I* measure found 13/26 variables to be significant at the 99% level ($p < 0.01$), 1/26 significant at the 95% level ($p < 0.05$), and 1/26 significant at the 90% level ($p < 0.10$). Lake access and game fish abundance was tested using join count statistics found statistically significant spatial autocorrelation results for most categories. For these variables, it is accepted that there is a degree of clustering or dispersion, see Table 2 and Appendix D. All climate variables and landscape metrics did not show spatial autocorrelation while most other variables did.

Table 2. Global Moran's I Level of Spatial Autocorrelation and Join Count Statistic Significance Values for all variables. Calculated using inverse distance weighting and Euclidean distance.

Lake Traits	Climate	Land cover composition	Land cover metrics
Perimeter	Mean temperature	Developed, open space***	AI, DO***
Elevation***	Temperature Range	Developed, medium	AI, EF***
Maximum Depth*	Mean precipitation	Deciduous forest***	ENN_AM,DO***
Access□			
Game Fish ı	Precipitation range	Mixed forest***	ENN_AM,EF***
Nearest distance to invaded lake		Scrub and shrubland***	RPR
Eurasian watermilfoil***			
Curly-leaf pondweed***			
Distance to i87 highway***		Emergent herbaceous wetland	SHEI
Distance to populated place**			
Presence-absence		Woody wetland***	
Eurasian watermilfoil***			
Curly-leaf pondweed***		Open water	

*Denotes < 10%, **Denotes < 5%, ***Denotes < 1% chance spatial pattern is random.
□ Access by carry down: $p < 0.01$, Access by public boat launch: $p > 0.1$.
ı No game fish: $p < 0.05$, One game fish: $p > 0.1$, Two game fish: $p < 0.05$, Three game fish: $p < 0.05$.

The join count statistics for EWM show that there are more absence/absence joins (actual = 32.12) than would be expected (27.22), and there are more presence/presence joins (actual = 11.95) than expected (7.22). The result is statistically significant at the 99% level for both, providing evidence for spatial autocorrelation. The join count statistics for CLP show that there are more absence/absence joins (actual = 51.62) than would be expected (48.84), and there are more presence/presence joins (actual = 2.29) than would

be expected (0.84). Both 0-0 and 1-1 joins are statistically significant at the 99% level, thus there is evidence of spatial autocorrelation. Permutation tests, not shown here, also had statistically significant results.

Table 3. Join-count statistics using two-sided test for presence-absence data. Calculated using inverse distance weighting.

Variable	Observed join		Expected join		Standard Deviate	
	0	1	0	1	0	1
Presence-absence of EWI	32.12	11.95	27.22	7.22	7.21***	8.88***
Presence-absence of CLP	51.62	2.29	48.84	0.84	5.56***	6.55***

0 = absent, 1 = present
*Denotes < 10%, **Denotes < 5%, ***Denotes < 1% chance spatial pattern is random.

The results for the local Moran's I autocorrelation can be seen in Figures 9 and 10 with significant areas ($p<0.05$) color coded according to the type of autocorrelation. The EWM presence-absence local autocorrelation shows spatial clusters of high values in the north and east, and clusters of low values in the midwest. There are only two outliers Fulton Chain Second and Fourth lakes (high – low values) and Green Pond (low – high value). For the CLP presence-absence local autocorrelation, there is less autocorrelation across the study region than for EWM presence-absence. There is only a cluster of high values and those are seen for southeast lakes and a chain of lakes in the north (Lower and Upper Chateaugay Lakes and Raquette Pond). Four outliers with high – low values are seen in the northern region (Lower Saranac Lake, Franklin Falls Flow, East Caroga Lake, Raquette Lake).

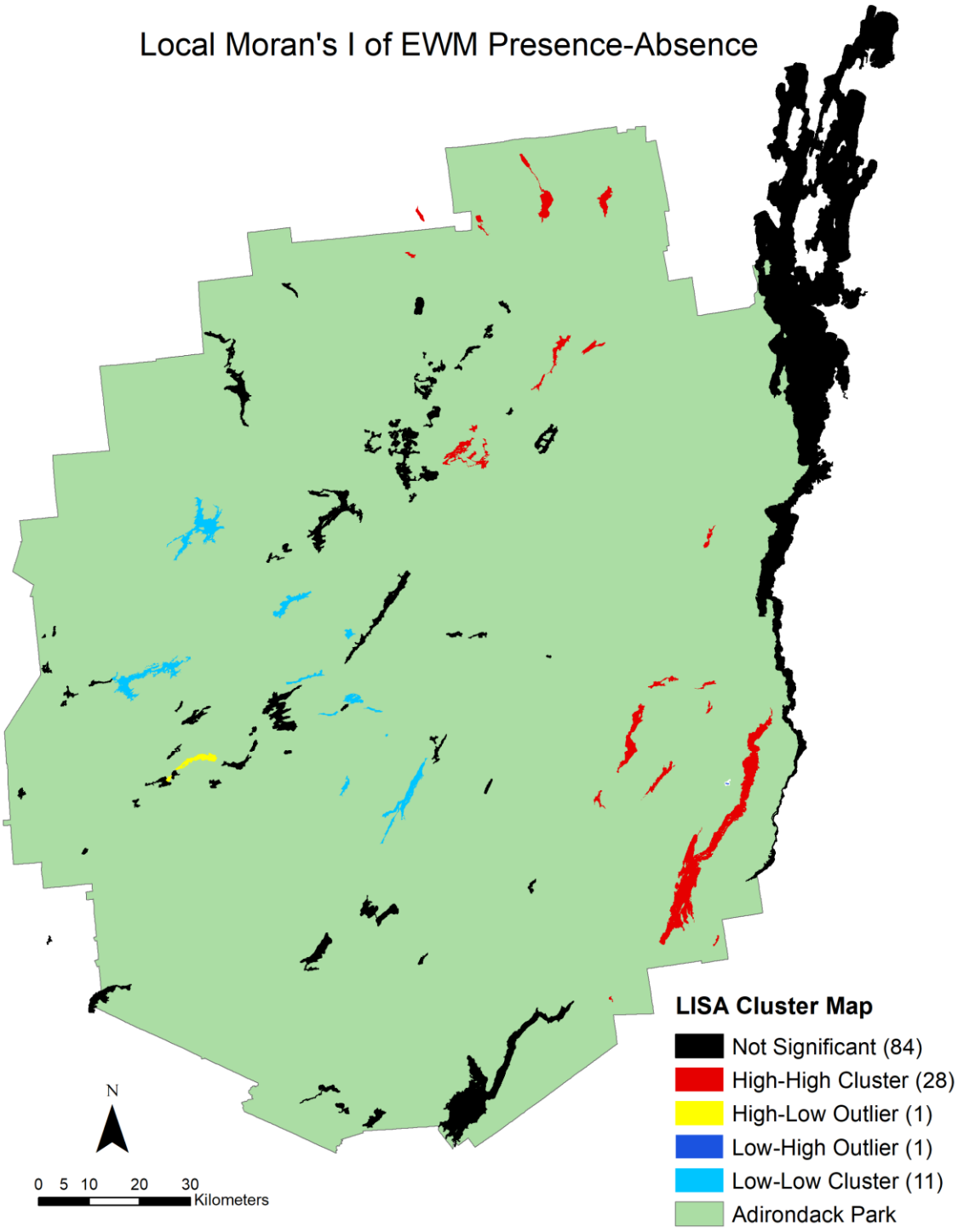


Figure 9. Local Moran's I of EWM Presence-Absence calculated by Inverse Distance Weighting.

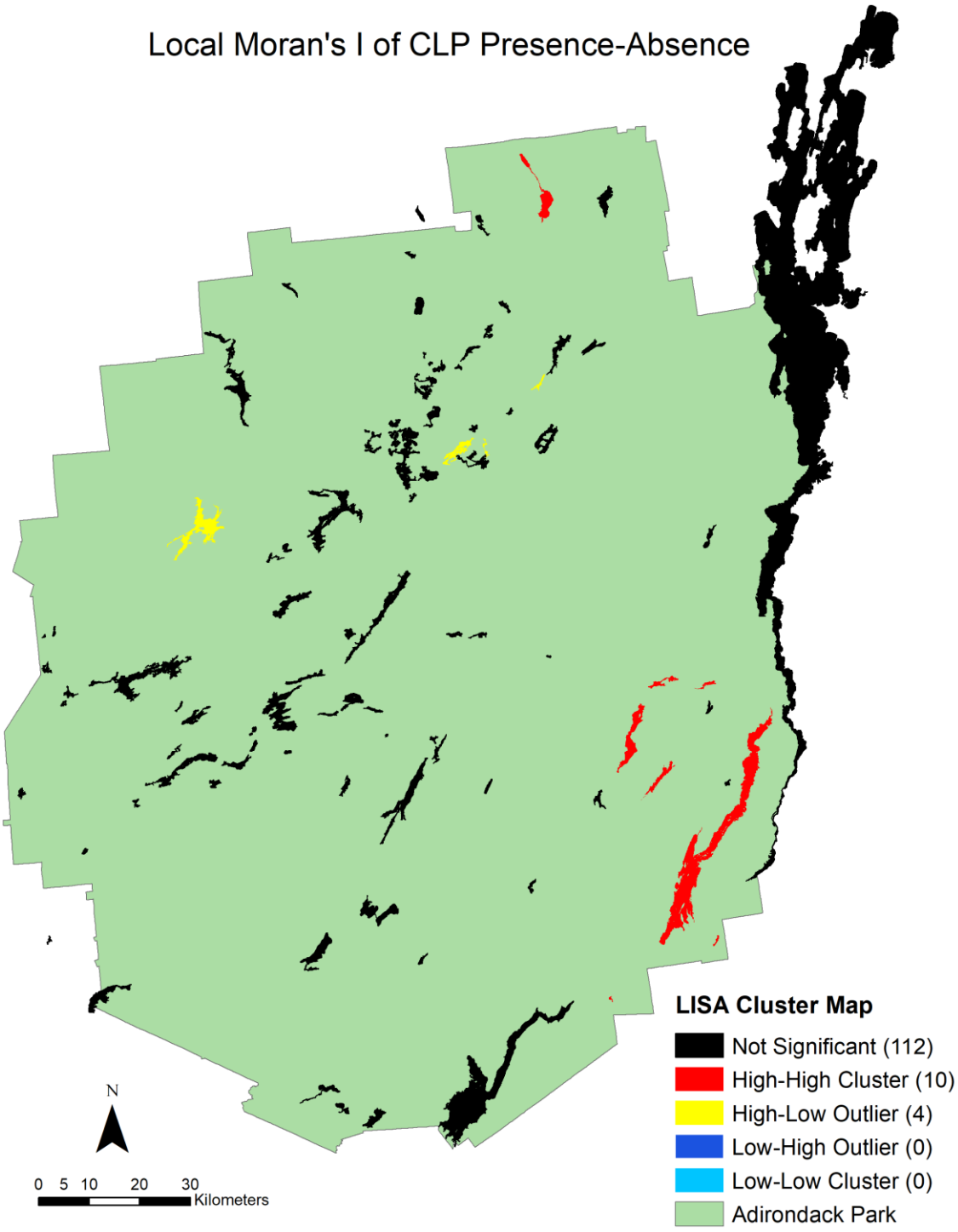


Figure 10. Local Moran's I of CLP Presence-Absence calculated by Inverse Distance Weighting.

4.2 Eurasian Watermilfoil Results

Nineteen predictor variables had statistically significant bivariate relationships with EWM presence-absence (Table 4). Among these, nearest distance to interstate highway I-87 exit had the highest pseudo R_L^2 value (R_L^2 = 0.19), followed by distance to nearest invaded lake (R_L^2= 0.18), percent deciduous forest (R_L^2 = 0.16), lake surface elevation (R_L^2 = 0.13), and percent developed open space (R_L^2 = 0.11). The remaining pseudo R_L^2 values were less than 0.10. The AICc values ranged from 134.39 for distance to nearest highway I-87 exit to 161.68 for SHEI. Extremely high odds ratios are seen for percent developed open space, percent developed medium intensity and percent mixed forest which may indicate overfitting. The four strongest negative associations to EWM presence include distance to nearest invaded lake (B = -2.80), distance to nearest I-87 highway exit (B = -2.79), percent deciduous forest (B = -2.54), and lake surface elevation (B = -2.05). The four strongest positive associations to EWM presence include percent developed open space (B = 1.76), AI of developed open space (B = 1.72), percent developed medium intensity (B = 1.52), and game fish abundance (B = 1.26).

These 19 variables were entered into the IT-AIC model selection approach. After an exhaustive method selection process evaluating 45300 models, the best model had AICc equal 81.90 with predictors lake surface elevation, nearest distance to nearest invaded lake, nearest distance to interstate highway exit I-87, percent developed open space, percent mixed forest, and SHEI. Five models were chosen amongst the top 100 model rankings for the LR and GWLR, using twelve predictors across the models. Two models were chosen from the top models that were within two AICc units of one another, indicating no difference in model performance, and others were chosen based on non-redundant variables to better explain EWM presence-absence. No final models had percent developed open space, AI of developed open space, and ENN of developed open space together in order to avoid multicollinearity issues ($R_{s,\ devo+aido}$ = 0.688, $R_{s,devo+endo}$ = 0.552, $R_{s,aido+endo}$ = 0.714). Percent developed open space and percent developed medium space were accepted for use in the same model because although they both represent human presence on land, they are essentially different: percent developed open space is representative of lawn grasses with < 20% and percent developed medium intensity is a mixture with a higher amount of impervious surfaces (50 – 79%).

The LR results can be seen in Table 5. Multicollinearity is of no concern in the models as all VIF values are below 2.5. The pseudo- R_L^2 values found the models to explain from 44% to 59% variation in EWM presence-absence. The AICc indicates that the models, from best to worst fit, are one, five, four, three, and two. Predictive success ranges from 85.7% for Model 4 to 88.1% for Model 1. No spatial autocorrelation is present in the model residuals, except for Model 2 with a Global Moran's *I* of 0.067 ($p < 0.05$), which implies that most models are not violating LR assumptions.

Table 4.

Bivariate logistic regression results for independent variables showing significance ($p < 0.10$) with presence-absence of EWM in Adirondack Park lakes as the response variable (N = 126).

Diagnostics include standardized beta values, Student's t ratio, and odds ratio. Goodness-of-fit diagnostics include Akaike's Information Criterion (corrected), chi square, and McFadden's R-square.
-- No relation observed. Independent model variables have been transformed to approach a normal distribution.

Independent variables	Standardized β	t ratio	Significance	Chi-Square	Significance	R-Square	AICc	Odds Ratio (C.I.)
Lake characteristics								
Perimeter	--	--	--	--	--	--	--	--
Maximum depth	--	--	--	--	--	--	--	--
Lake surface elevation	-2.046	-3.812	<0.001	20.788	<0.001	0.1285	145.062	0.889 (0.844 - 0.937)
Distance to nearest invaded lake	-2.800	-4.382	<0.001	29.956	<0.001	0.1823	136.369	0.889 (0.844 - 0.937)
Distance to nearest I-87 highway exit	-2.794	-4.567	<0.001	31.461	<0.001	0.1945	134.389	0.951 (0.931 - 0.972)
Distance to nearest populated place	--	--	--	--	--	--	--	--
Access type	0.924	2.175	0.030	5.025	0.025	0.0311	160.825	2.457 (1.093 - 5.521)
Game fish abundance	1.261	2.932	0.003	9.329	0.002	0.0577	156.521	1.951 (1.248 - 3.049)
Climate								
Mean temperature	1.121	2.703	0.007	7.992	0.005	0.0494	157.858	<0.001 (<0.001 - 0.054)
Temperature range (max - min)	--	--	--	--	--	--	--	--
Mean precipitation	-0.877	-2.032	0.042	4.46	0.035	0.0276	161.390	0.868 (0.757 - 0.995)
Precipitation range (max - min)	0.828	2.051	0.040	4.438	0.035	0.0274	161.412	1.971 (1.031 - 3.768)
Land cover & class metrics†								
Percent developed, open space	1.761	3.768	<0.001	17.118	<0.001	0.1058	148.732	1871.606 (37.17 -
Percent developed, medium intensity	1.518	3.294	<0.001	13.239	<0.001	0.0818	152.611	8*10^7 (1656.12 -
Percent deciduous forest	-2.535	-4.382	<0.001	26.713	<0.001	0.1651	139.137	0.912 (0.875 - 0.95)
Percent mixed forest	0.967	2.321	0.020	5.801	0.016	0.0359	160.049	124.781 (2.119 - 7348.849)
Percent shrub and scrubland	-1.403	-2.714	0.007	9.064	0.003	0.056	156.786	<0.001 (<0.001 - 0.118)
Percent woody wetland	--	--	--	--	--	--	--	--
Percent emergent herbaceous wetland	--	--	--	--	--	--	--	--
Percent open water	0.995	2.413	0.016	6.145	0.013	0.038	159.705	1.04 (1.007 - 1.074)
Aggregation index of developed, open space	1.721	3.026	0.002	12.6	<0.001	0.078	153.231	1.035 (1.012 - 1.058)
Aggregation index of evergreen forest	--	--	--	--	--	--	--	--
Euclidean nearest neighbor of developed, open space	0.922	2.041	0.041	4.657	0.031	0.0288	161.193	1.625 (1.02 - 2.591)
Euclidean nearest neighbor of evergreen forest	-0.94	-2.042	0.041	4.799	0.028	0.0297	161.050	0.105 (0.012 - 0.913)
Landscape diversity†								
Relative patch richness (RPR)	1.129	2.671	0.008	7.665	0.006	0.0474	158.185	1.037 (1.01 - 1.065)
Shannon's evenness index (SHEI)	-0.717	-1.747	0.081	3.169	0.075	0.0196	162.681	0.04 (0.001 - 1.479)

Symbol designation: † Associated variables calculated using a 300-meter riparian zone landscape for each lake.

Table 5.

Logistic regression modeling results for the prediction of presence-absence of EWM across 126 lakes within the Adirondack Region of New York for 2016.
-- No relation observed. Covariate values are standardized beta values; values enclosed in parentheses are individual *p* -values of the t statistic. Independent model variables have been transformed to approach a normal distribution and standardized to a mean of 0 and variance of 1.

Statistical measures and independent variables	Models				
Statistical measures	Model 1	Model 2	Model 3	Model 4	Model 5
Akaike's Information Criterion, corrected	81.914	90.380	86.405	85.310	83.398
McFadden's Pseudo R-Square	0.586	0.520	0.442	0.551	0.577
Percent correctly predicted (0.5 cut off)	88.10%	86.51%	86.51%	85.71%	88.89%
Residual Global Moran's *I*	-0.012	0.067**	-0.011	0.024	-0.023
Variance Inflation Factor (VIF) max value	1.457	1.280	1.404	1.392	1.812
Independent variables					
Logistic regression beta constant	0.000 (<0.001)	0.000 (<0.001)	0.000 (<0.001)	0.000 (<0.001)	0.000 (<0.001)
Lake characteristics					
Lake surface elevation	-1.979 (0.013)	-2.426 (0.002)	-2.369 (0.003)	-1.647 (0.032)	--
Distance to nearest invaded lake	-5.764 (<0.001)	-4.996 (<0.001)	-5.097 (<0.001)	-5.725 (<0.001)	-5.338 (<0.001)
Distance to nearest interstate highway exit (I-87)	-2.834 (0.026)	--	-2.763 (0.025)	-4.223 (<0.001)	-4.678 (<0.001)
Game fish abundance	--	--	1.254 (0.060)	--	--
Climate					
Mean precipitation	--	-1.351 (0.043)	--	--	--
Land cover & class metrics†					
Percent developed, open space	2.442 (0.003)	1.564 (0.029)	--	1.79 (0.013)	1.690 (0.052)
Percent developed, medium intensity	--	--	--	--	1.586 (0.051)
Percent deciduous forest	--	-1.913 (0.014)	--	--	--
Percent mixed forest	1.977 (0.020)	--	1.396 (0.075)	--	1.364 (0.060)
Percent open water	--	--	--	1.522 (0.036)	1.887 (0.011)
Aggregation index of developed, open space	--	--	2.055 (0.019)	--	--
Landscape diversity†					
Shannon's evenness index (SHEI)	-2.559 (0.011)	--	--	--	--

Symbol designation: † Associated variables calculated using a 300-meter riparian zone landscape for each lake. ** Significant at the 95% level.

All twelve predictors were significant in the models, indicating that they are good predictors for EWM presence-absence in multivariate models. The IT-AICc selection process averaged the most important variables by their appearance in the top 100 models (Figure 11).

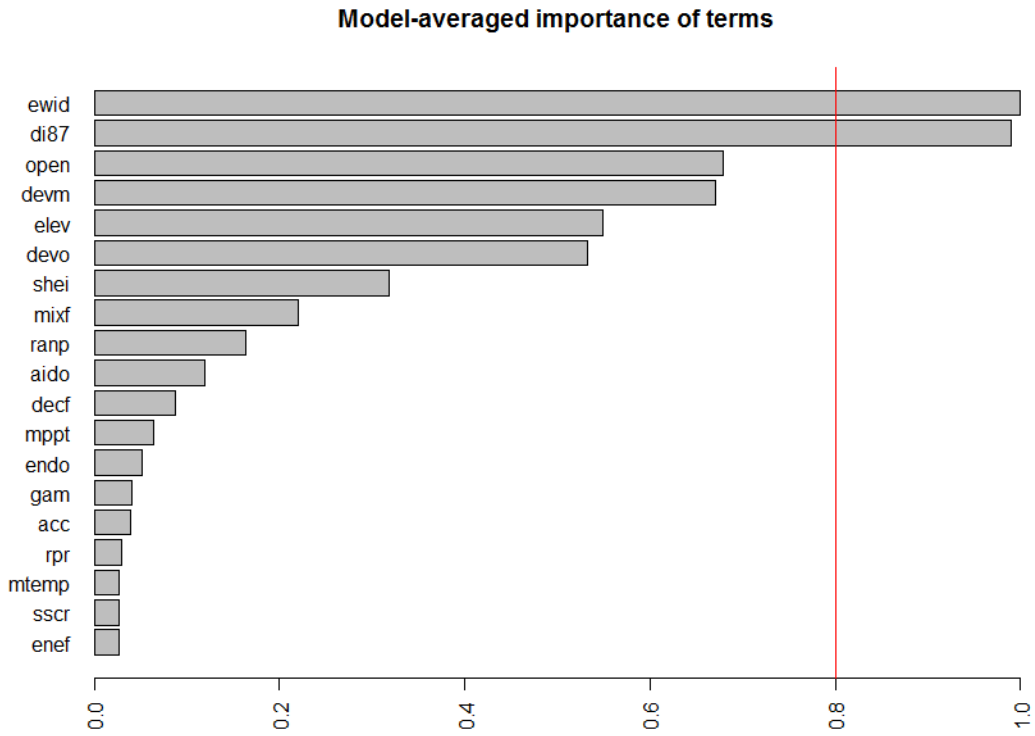


Figure 11. Importance of terms based on number of appearances in top 100 models chosen in LR multi-model selection process for the prediction of EWM presence-absence with a maximum of 6 predictors. The red line marks 80%.

Taking into account the bivariate and multivariate regression results, the two most important predictors are distance to nearest invaded lake and distance to nearest I-87 exit as they appear in more than 80% of the top 100 models, contribute to most models selected for the study at the 99% and 95% level, and have the strongest pseudo $R_L^2$and beta values. Between the two, distance to nearest invaded lake is likely to be slightly more important as it is significant at the 99% level in every model, and exceeds appearances in the top 100 models (Figure 11). The next two important predictors include percent developed open and lake surface elevation as they are in 4 models and significant at the 95% level. Percent mixed forest in the riparian zone is another important predictor, as it is in 3 models, significant at the 90% level, and is a predictor in the top models within 2 AICc units of one

another selected by the IT-AICc method. The remaining predictors are less important but still significantly contribute at the 90 – 90% level, including SHEI, percent open water, percent deciduous forest, percent developed medium intensity, mean precipitation, and AI of developed open space. Furthermore, the predictors association with EWM presence-absence across Adirondack Park is negative for lake surface elevation, distance to nearest invaded lake, distance to nearest I-87 highway exit, mean precipitation, and SHEI; the association is positive for game fish abundance, human presence landscape predictors of developed open space and medium intensity, mixed forest, and open water.

The GWLR model results are shown in Table 6. An adaptive Gaussian kernel based on number of neighbors was used because the data points in the study area are irregularly positioned. The minimum range of neighbors was set to thirty to warrant overfitting from occurring. Of all the models, the number of neighbors used ranged from 28.21% (in Model 4) to 74.03% (in Model 5). From the twelve predictors used in the models, those that exhibit non-stationarity based on the previous Global Moran's *I* results include lake surface elevation, game fish abundance, distance to nearest invaded lake, distance to nearest I-87 highway exit, percent developed open space, percent deciduous forest, AI of developed open space, and the response variable EWM presence-absence. The other five independent variables are stationary. The models ranked from best to worst based on AICc results are: Model 4 (AICc = 78.70), Model 1 (AICc = 79.56), Model 5 (AICc = 82.65), Model 2 (AICc = 84.24), and Model 3 (85.89). Between 60% to 75% of EWM presence-absence variation was explained by the GWLR models. The models successfully predicted between 87% to 91% of EWM presence-absence.

The GWLR model diagnostics show improvements based on goodness of fit and predictive success relative to the LR models (see Appendix F for model comparison table). The explained variation of EWM presence-absence increased by an average of 23% (1.9 – 33%), and the predictive success increased by 2.6%. Other metrics used to compare the two included deviance and RMSE, of which both were shown to decrease which indicates model improvement. Spatial autocorrelation was not present in any residuals for GWLR, showing an improvement for Model 2 that exhibited spatial autocorrelation with the LR model. Also, the global relationships to EWM presence-absence with the predictors held true at a local level with the same positive or negative standardized betas.

Table 6.

Geographically weighted regression modeling results for the prediction of presence-absence of EWM across 126 lakes within the Adirondack Region -- No relation observed. Independent model variables have been transformed, and standardized to set the mean at 0 and variance to 1.

Statistical measures and independent variables	Models				
Statistical Measures	Model 1	Model 2	Model 3	Model 4	Model 5
Akaike's Information Criterion (AICc)	79.560	84.240	85.891	78.698	82.648
Adaptive Kernel Neighbors	50.79%	34.69%	69.59%	28.21%	74.03%
Number Parameters	10.051	12.003	8.909	12.977	8.821
McFadden's Pseudo R-Square	0.597	0.645	0.589	0.694	0.751
Percent correctly predicted (0.5 cut off)	89.68%	90.48%	87.30%	89.68%	89.68%
Residual Global Moran's *I*	-0.026	0.018	-0.015	-0.050	-0.025
Local Regression Parameter Descriptive Statistics: (Median)					
Constant	-2.761	-2.341	-2.341	-3.303	-2.573
Lake characteristics					
Lake surface elevation	-0.890	-0.825	-1.041	-0.826	--
Distance to nearest invaded lake	-3.229	-2.659	-2.650	-3.634	-2.739
Distance to nearest interstate highway exit (I-87)	-1.512	--	-1.349	-2.337	-2.233
Game fish abundance	--	--	0.581	--	--
Climate					
Mean precipitation	--	-0.651	--	--	--
Land cover & class metrics†					
Percent developed, open space	1.085	0.663	--	0.762	0.743
Percent developed, medium intensity	--	--	--	--	0.740
Percent deciduous forest	--	-1.064	--	--	--
Percent mixed forest	0.983	--	0.681	--	0.703
Percent open water	--	--	--	1.110	0.958
Aggregation index of developed, open space	--	--	0.968	--	--
Landscape diversity †					
Shannon's evenness index (SHEI)	-1.381	--	--	--	--

Symbol designations: † Associated variables calculated using a 300-meter riparian zone landscape for each lake;

To gain a greater understanding of local variations among model parameters, the estimates from Model 1 were mapped, shown on the proceeding pages. Model 1 was chosen because it consistently had low AICc values and its predictors are of highest importance. The map of distance to nearest interstate highway I-87 exit (Figure 12) shows a negative association in the northwest region and decreasing southward except a large portion of the park in the south has no statistical association to the predictor. Distance to nearest invaded lake has the strongest association to EWM presence-absence and the map (Figure 13) shows this association to be strongest in the northwest (from Stony Creek Pond, through Saranac Lakes up to Loon Lake) region and weakest in the south of the park. Lake surface elevation has a weaker negative association, and its local variation shows there to be no statistical significance in the mid-western part of the park and the strongest association in the west (Figure 14). Developed open space in the riparian zone, has strong associations in the south decreasing towards the north (quite weak in the northwest) with a region in the northwest of no association (Figure 15). Mixed forest in the riparian zone has a positive association to EWM presence and is strongest in the middle of the park, weakening towards the north and Lake Champlain with a small part in the north having no statistical association (Figure 16). Lastly, SHEI is negatively and significantly associated with EWM presence-absence, with the strongest estimated parameters in the northwest and weakest in the south (Figure 17). Furthermore, to understand best where Model 1 parameters predict EWM presence-absence, the local R_L^2 values were mapped (Figure 18). The map shows that Model 1 best applies in the east of Adirondack Park with 67% explained variation and is weaker in western parts with 60% explained variation.

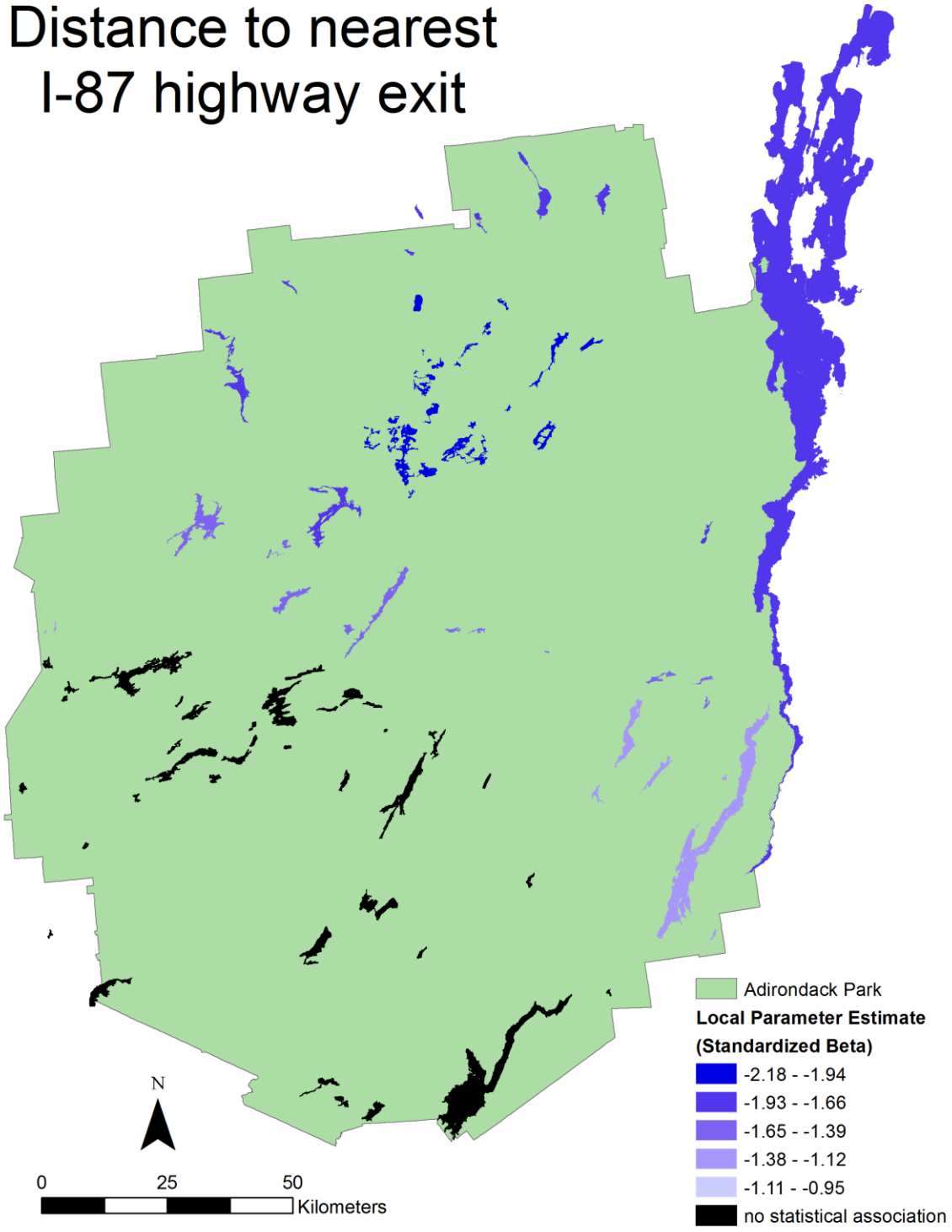


Figure 12. Parameter estimate of distance to nearest highway I-87 exit from EWM presence-absence GWLR Model 1.

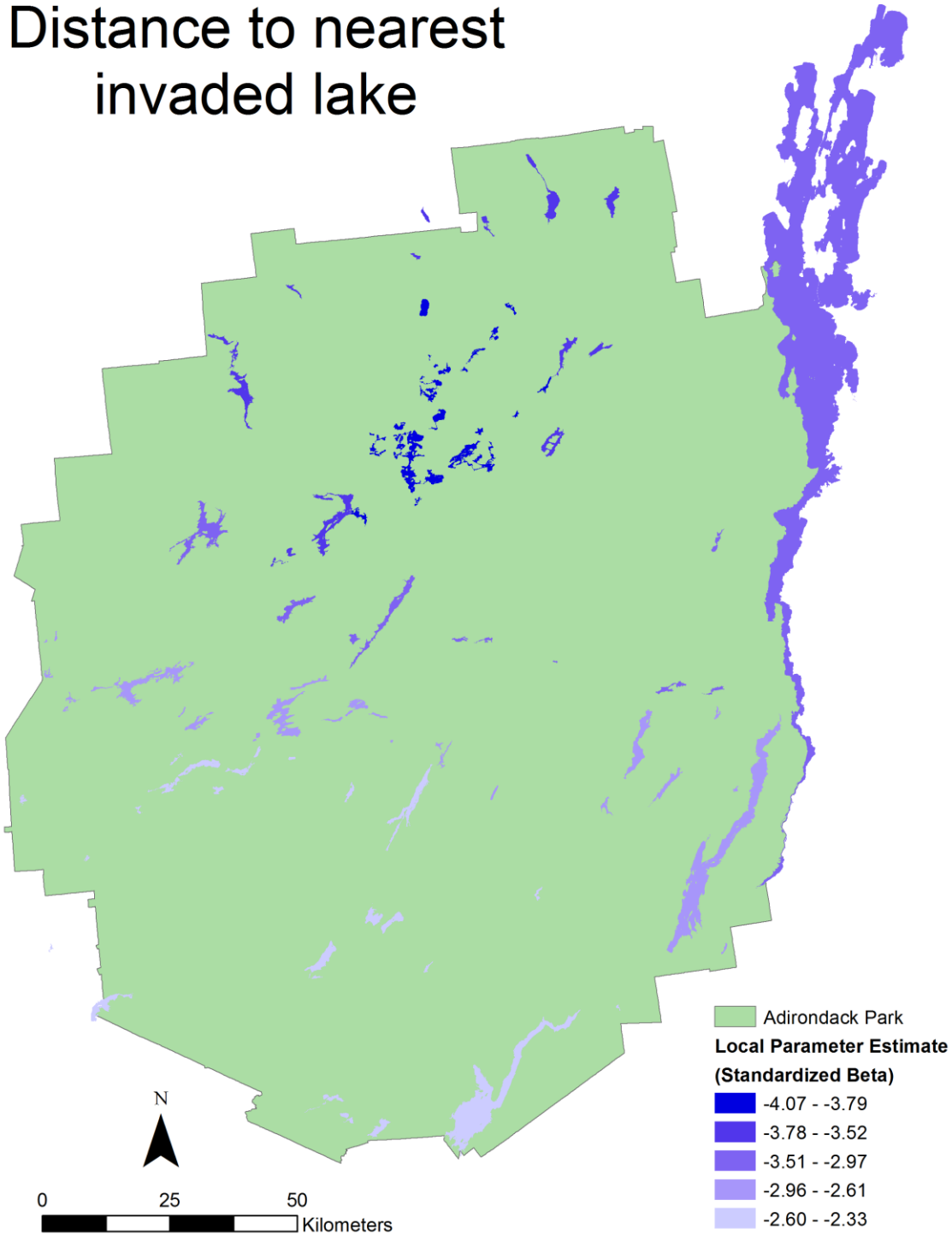


Figure 13. Parameter estimate of distance to nearest invaded lake from EWM presence-absence GWLR Model 1.

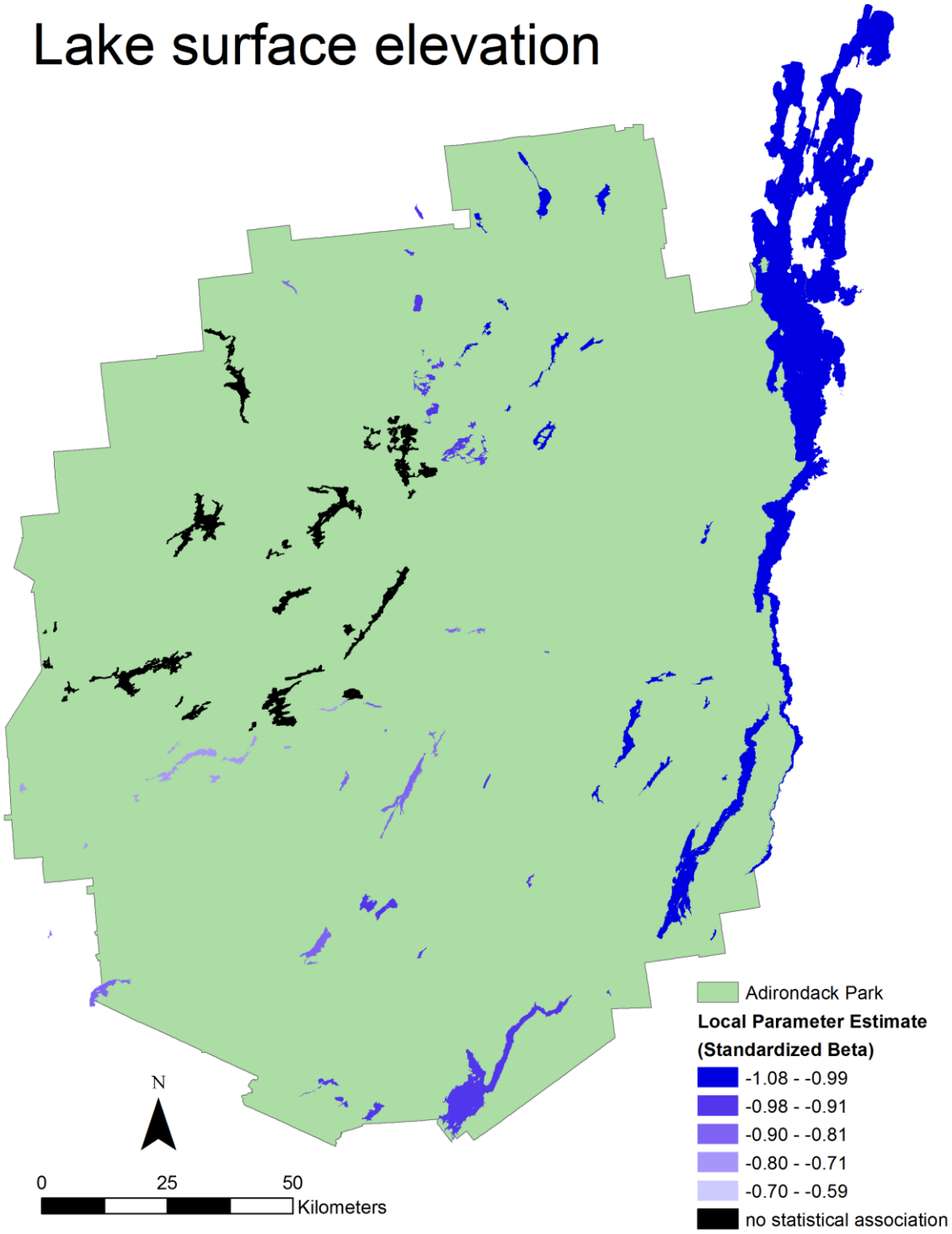


Figure 14. Parameter estimate of lake surface elevation from EWM presence-absence GWLR Model 1.

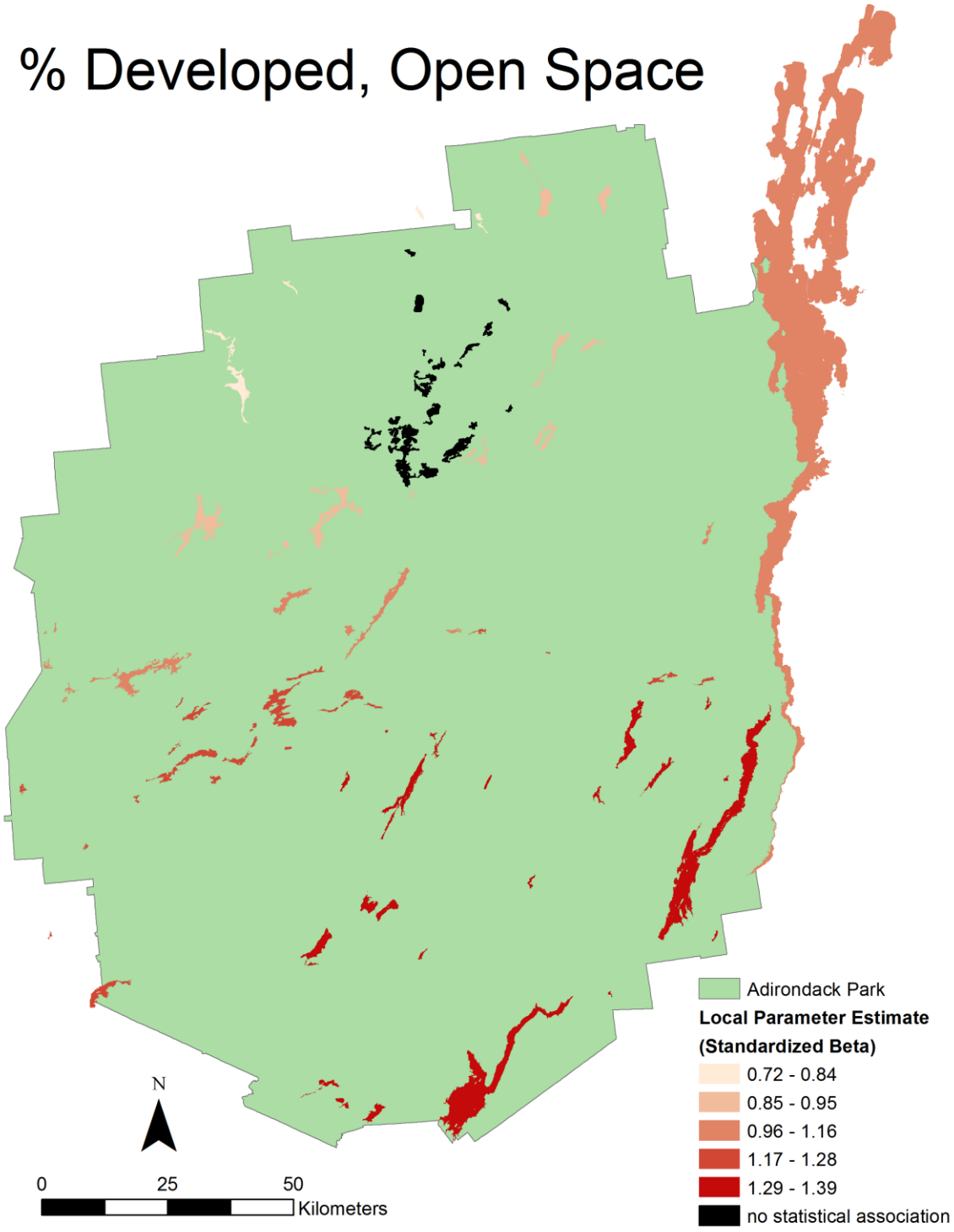


Figure 15. Parameter estimate of percent developed, open space in the riparian zone (300meter buffer of lakes) from EWM presence-absence GWLR Model 1.

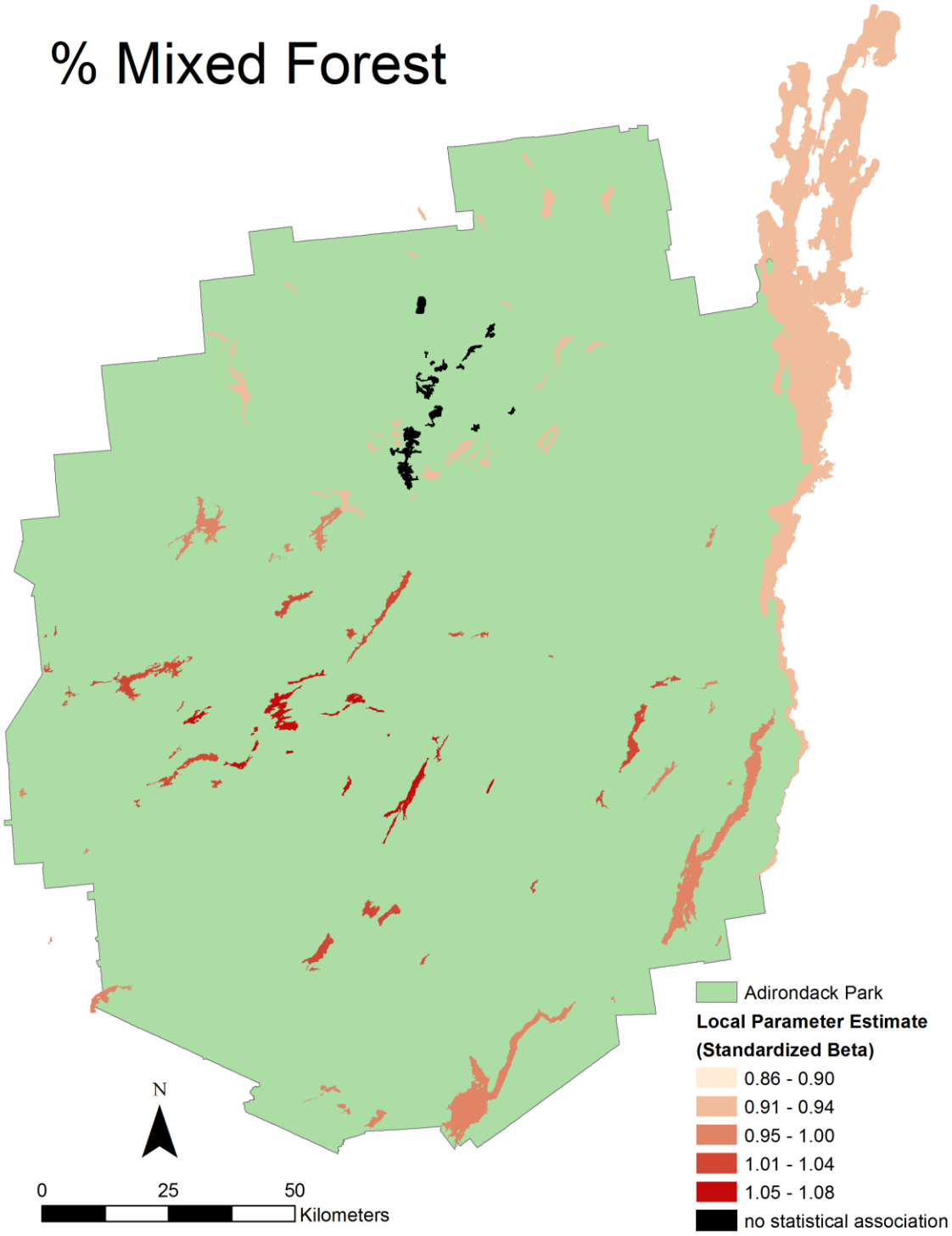


Figure 16. Parameter estimate of percent mixed forest in the riparian zone (300meter buffer of lakes) from EWM presence-absence GWLR Model 1.

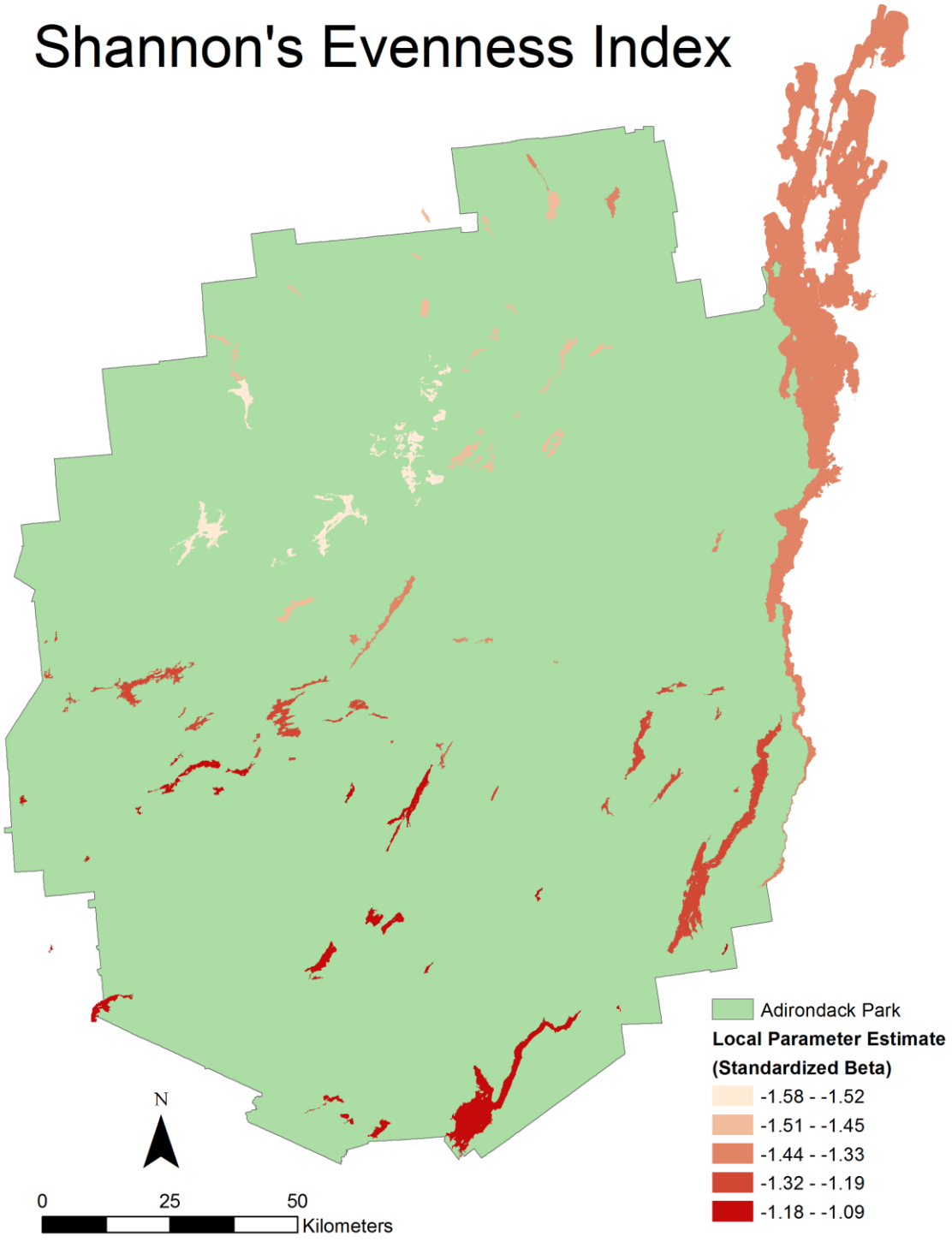


Figure 17. Parameter estimate of Shannon's Evenness Index in the riparian zone (300meter buffer of lakes) from EWM presence-absence GWLR Model 1.

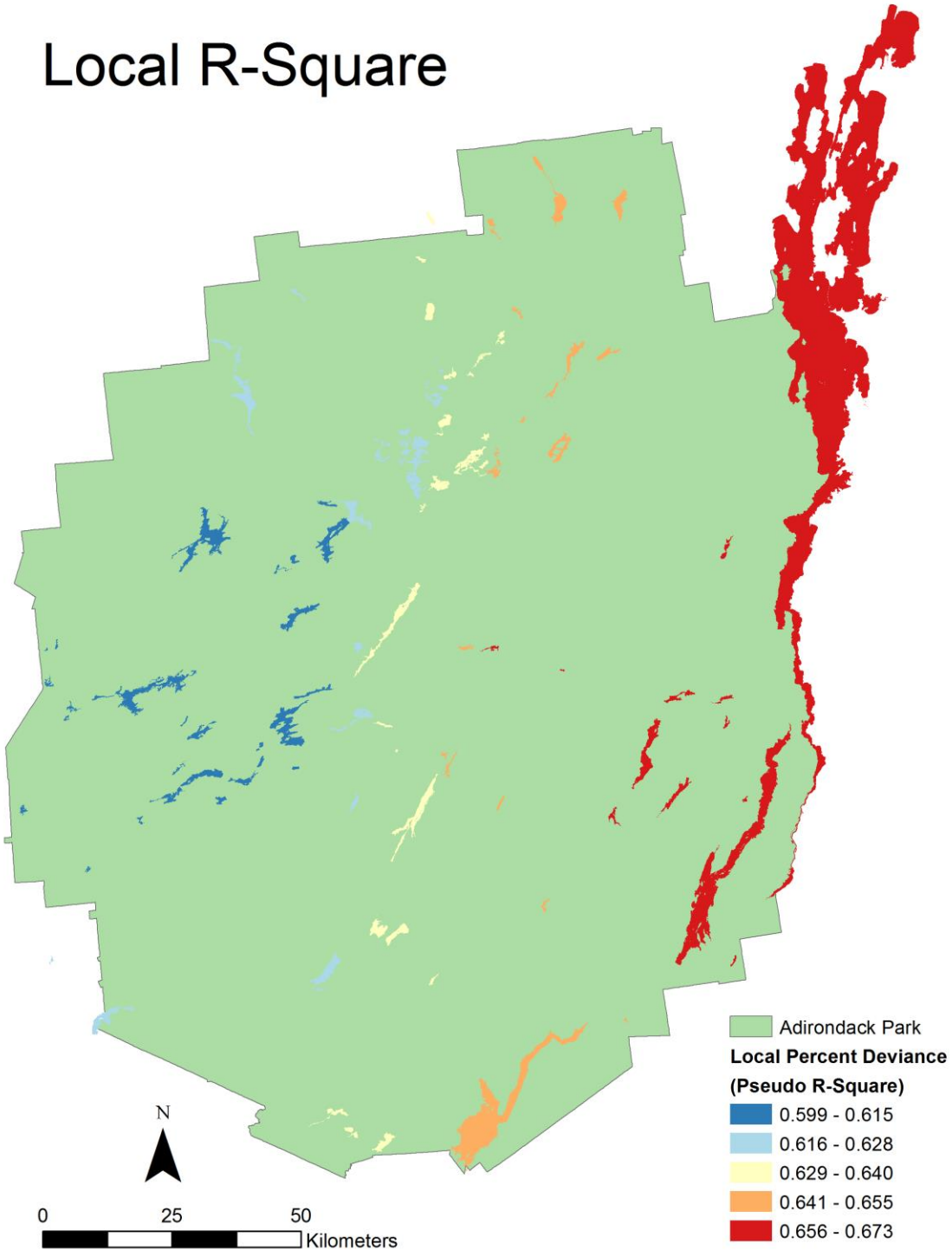


Figure 18. Local percent deviance from EWM presence-absence GWLR Model 1.

4.3 Curly leaf pondweed Results

The bivariate logistic regression for variables in the CLP presence-absence analysis found 13 variables to be statistically significant predictors (Table 7). The variable with the highest pseudo R_L^2 was distance to nearest interstate highway I-87 (R_L^2 = 0.30), followed by lake surface elevation (R_L^2 = 0.29), game fish abundance (R_L^2 = 0.20), distance to nearest invaded lake (R_L^2 = 0.16), relative patch richness (R_L^2 = 0.15), percent deciduous forest (R_L^2 = 0.11), percent developed open space (R_L^2 = 0.11) and the remaining variables had values less than 0.10. The AICc values range from 68.258 for distance to nearest interstate highway I-87 to 93.271 for percent open water. Extremely high odds ratios are seen for percent developed open space, percent developed medium intensity and mean temperature. Fairly strong negative associations to CLP presence-absence at the global level are seen with predictors distance to nearest I-87 exit (B = -5.85), distance to nearest invaded lake (B = -4.70), lake surface elevation (B = -4.01), and percent deciduous forest (B = -3.27). The strongest positive association is with game fish abundance (B = 4.38), with the second strongest being RPR (B = 3.38).

All thirteen predictors were entered into the IT-AICc multi-model selection and by an exhaustive method approach, the top 100 models were ranked by AICc and five models chosen to best elucidate results. The LR model results are shown in Table 8. There are no multicollinearity issues as all VIF values are less than 2.5. Pseudo R_L^2 values indicate that explained variation of CLP presence-absence ranges from 47% to 50%. Models 1, 2, and 4 are within two AICc units of one another and so they are considered to be equivalent in terms of model fit and all are the "best" models. Model 3 and 5 have higher AICc units and so are not considered among the best models for LR. The prediction success was quite similar for all models ranging from 92% to 94%. There was no spatial autocorrelation present in the residuals providing evidence that the errors were independent and LR model assumptions met.

Furthermore, nine of the original thirteen predictors were included in the models although not all were significant. Those that were consistently significant include lake surface elevation, distance to nearest invaded lake, and game fish abundance and so are good predictors for CLP presence-absence in global multivariate models. The most

Table 7.

Bivariate logistic regression results for independent variables showing significance ($p < 0.10$) with presence-absence of CLP in Adirondack Park lakes as the response variable (N = 126). Diagnostics include standardized coefficients (beta values), Student's t ratio, and odds ratio. Goodness-of-fit diagnostics include Akaike's Information Criterion (corrected), chi-square and McFadden's pseudo R-square. Independent model variables have been transformed to approach a normal distribution.

Independent variables	Standardized	t ratio	Significance	Chi-Square	Significance	R-Square	AICc	Odds Ratio (C.I.)
Lake characteristics								
Perimeter	2.354	2.896	0.004	9.909	0.003	0.099	86.989	5.742 (1.759 - 18.74)
Maximum depth	--	--	--	--	--	--	--	-
Lake surface elevation	-4.005	-4.47	<0.001	27.524	<0.001	0.299	68.559	0.988 (0.982 - 0.993)
Distance to nearest invaded lake	-4.701	-2.998	0.003	14.791	<0.001	0.161	81.292	0.921 (0.873 - 0.972)
Distance to nearest interstate highway exit (I-	-5.845	-4.27	<0.001	27.826	<0.001	0.303	68.258	0.931 (0.901 - 0.962)
Game fish abundance	4.383	3.642	<0.001	18.730	<0.001	0.204	77.353	4.884 (2.08 - 11.468)
Climate								
Mean temperature	1.382	1.959	0.050	3.510	0.061	0.038	92.555	11910.332 (0.996 - 1*10^8)
Temperature range (max - min)	--	--	--	--	--	--	--	-
Mean precipitation	--	--	--	--	--	--	--	-
Precipitation range (max - min)	--	--	--	--	--	--	--	-
Land cover & class metrics†								
Percent developed, open space	2.556	2.989	0.003	9.636	0.002	0.105	86.426	754.932 (13.092 - 235250.854)
Percent developed, medium intensity	1.938	2.768	0.006	7.567	0.006	0.082	88.514	8*10^6 (105.634 - 6*10^11)
Percent deciduous forest	-3.272	-2.74	0.006	9.816	0.002	0.107	86.257	0.922 (0.87 - 0.977)
Percent mixed forest	--	--	--	--	--	--	--	-
Percent shrub and scrubland	-2.835	-2.236	0.025	6.445	0.011	0.070	89.63	<0.001 (<0.001 - 0.269)
Percent woody wetland	--	--	--	--	--	--	--	-
Percent emergent herbaceous wetland	--	--	--	--	--	--	--	-
Percent open water	1.362	1.688	0.091	2.821	0.093	0.031	93.271	1.038 (0.994 - 1.083)
Aggregation index of developed, open space	--	--	--	--	--	--	--	-
Aggregation index of evergreen forest	--	--	--	--	--	--	--	-
Euclidean nearest neighbor of developed, open	--	--	--	--	--	--	--	-
Euclidean nearest neighbor of evergreen forest	--	--	--	--	--	--	--	-
Landscape diversity†								
Relative patch richness (RPR)	3.383	3.407	<0.001	14.136	<0.001	0.154	81.947	1.077 (1.032 - 1.124)
Shannon's evenness index (SHEI)	-2.196	-2.428	0.015	6.416	0.011	0.070	89.672	0.001 (<0.001 - 0.274)

Symbol designation: † Associated variables calculated using a 300-meter riparian zone landscape for each lake.

Table 8.

Logistic regression modeling results for the prediction of presence-absence of CLP across 126 lakes within the Adirondack Region of New York for 2016. -- No relation observed. Covariate values are standardized beta values; values enclosed in parentheses are individual *p* -values of the t statistic. Independent variables have been transformed to approach a normal distribution and standardized to a mean of 0 and variance of 1.

Statistical measures and independent variables	Models				
Statistical measures	Model 1	Model 2	Model 3	Model 4	Model 5
Akaike's Information Criterion, corrected	51.614	53.228	54.265	53.780	57.575
McFadden's Pseudo r-square	0.530	0.584	0.501	0.530	0.465
Percent correctly predicted (0.5 cut off)	94.44%	94.44%	92.06%	94.44%	92.06%
Residual Global Moran's I	-0.010	-0.011	0.000	-0.010	0.011
Variance Inflation Factor (VIF) max value	1.060	2.267	1.158	1.074	1.276
Independent variables					
Logistic regression beta constant	0 (<0.001)	0 (<0.001)	0 (<0.001)	0 (<0.001)	0 (<0.001)
Lake characteristics					
Perimeter	--	3.300 (0.155)	--	--	--
Lake surface elevation	-4.213 (<0.001)	-3.912 (0.002)	-3.174 (0.011)	-4.203 (<0.001)	-1.140 (0.001)
Distance to nearest invaded lake	-4.921 (0.043)	-5.608 (0.047)	--	-4.897 (0.046)	-2.047 (0.009)
Distance to nearest interstate highway exit (I-87)	--	--	-3.224 (0.059)	--	--
Game fish abundance	3.521 (0.008)	4.926 (0.020)	3.832 (0.007)	3.506 (0.009)	--
Climate					
Mean temperature	--	--	--	0.066 (0.95)	--
Land cover & class metrics†					
Percent shrub and scrubland	--	-2.057 (0.319)	--	--	--
Percent open water	--	-4.359 (0.071)	--	--	--
Landscape diversity†					--
Relative patch richness (RPR)	--	--	--	--	0.701 (0.093)

Symbol designation: † Associated variables calculated using a 300-meter riparian zone landscape for each lake.

important predictor is lake surface elevation as it is found in every model, significant at the 95% level, and had the second highest R_L^2 value. The second most important variables are game fish abundance and distance to nearest invaded lake, present in four models and significant at the 95% level. Distance to nearest I-87 highway exit was only present in one model, significant at the 90% level, even though it had the highest R_L^2 amongst the bivariate regressions. Weaker variables include percent open water and RPR in the riparian zone, only present in one model but both significant at the 90% level. Other predictors present in the models but not significant include perimeter, percent shrub and scrubland in the riparian zone, and mean temperature. These variables will not be discussed hereafter as contributors leaving six significant predictors of CLP presence-absence across Adirondack Park. Fairly strong associations are seen with the standardized beta values, with the top three predictors ranging from three to five absolute beta. Negative relationships to CLP presence-absence include lake surface elevation, distance to nearest invaded lake, distance to nearest I-87 highway exit, and percent open water. Positive relationships are only present with game fish abundance, and RPR.

Using an adaptive Gaussian kernel due to irregularly positioned study points, the GWLR model results were obtained and are shown in Table 9. The number of neighbors ranged from 17% in Model 3 to 100% in Model 2. Of the important variables, non-stationarity shown by Global Moran's *I* was found for lake surface elevation, game fish abundance, distance to nearest invaded lake, distance to I-87 highway exit, and the response variable CLP presence-absence. Percent open water and RPR were the only stationary variables. Furthermore, the best fitting GWR models include Model 1 (AICc = 50.20) and Model 4 (AICc = 52.93). The five models explain between 56% and 66% of the variation and successfully predict between 93% and 94% of CLP presence-absence. No residuals have significant Moran's I values, validating error independence. Relative to LR, the GWLR results show an improvement in goodness of fit, and reaffirms the global associations. There is an average 16% (4 – 32%) increase in explained variation of CLP presence-absence. Predictive succession rate only slightly increases with 0.8% improvement. Other diagnostic statistics, deviance and RMSE, both decrease indicating a better model fit.

Table 9.

Geographically weighted regression modeling results for the prediction of presence-absence of CLP across 126 lakes within the Adirondack Region
-- No relation observed. Independent model variables have been transformed and standardized to set the mean at 0 and variance to 1.

Statistical measures and independent variables	Model				
Diagnostic Statistics	Model 1	Model 2	Model 3	Model 4	Model 5
Akaike's Information Criterion (AICc)	50.202	53.117	55.617	52.927	54.178
Adaptive Kernel Neighbors	56.53%	100.00%	16.81%	59.98%	47.18%
Number Parameters	5.899	7.928	11.068	7.106	6.435
McFadden's Pseudo R-square	0.590	0.608	0.662	0.590	0.560
Percent correctly predicted (0.5 cut off)	94.44%	94.44%	93.65%	94.44%	94.44%
Residuals Global Moran's *I*	-0.014	-0.011	-0.014	-0.012	-0.026
Local Regression Parameter Descriptive Statistics: (Median)					
Constant	-3.745	-4.532	-4.167	-3.713	-3.748
Lake characteristics					
Perimeter	--	1.127	--	--	--
Lake surface elevation	-1.257	-1.251	-1.200	-1.258	-0.864
Distance to nearest invaded lake	-1.268	-1.753	--	-1.237	-1.754
Distance to nearest interstate highway exit (I-87)	--	--	-1.378	--	--
Game fish abundance	1.250019	1.587	1.141	1.232	--
Climate					
Mean temperature	--	--	--	0.0971	--
Land cover & class metrics					
Percent shrub and scrubland	--	-0.733	--	--	--
Percent open water	--	-1.457	--	--	--
Landscape diversity †					
Relative patch richness (RPR)	--	--	--	--	0.703

Symbol designations: † Associated variables calculated using a 300-meter riparian zone landscape for each lake.

The estimated parameters from Model 1, as it is the best fitting model with significant predictors, to explore local variation. The map of lake surface elevation shows an association across the entire region with the strongest negative association in the east and the lowest in the northwest, which is expected because of elevation gradients (Figure 19). Game fish abundance has the strongest positive association (up to 1.62) in a few lakes in the northwest and is weakest (down to 0.97) in the southeast region (Figure 20). The majority of the region has no statistical association between CLP presence-absence and distance to nearest invaded lake, seen in the map in the mid and west parts of Adirondack Park. Lake Champlain, Lake George and nearby lakes like Paradox Lake have the strongest negative associations between -1.77 and -1.99 (Figure 21). Great Sacandaga Lake to the south of Lake Champlain has a slightly weaker association then these. Over the entire region, the model best performs in the southeast where explained variation reaches 61% and is much weaker in the northwest with 41% (Figure 22).

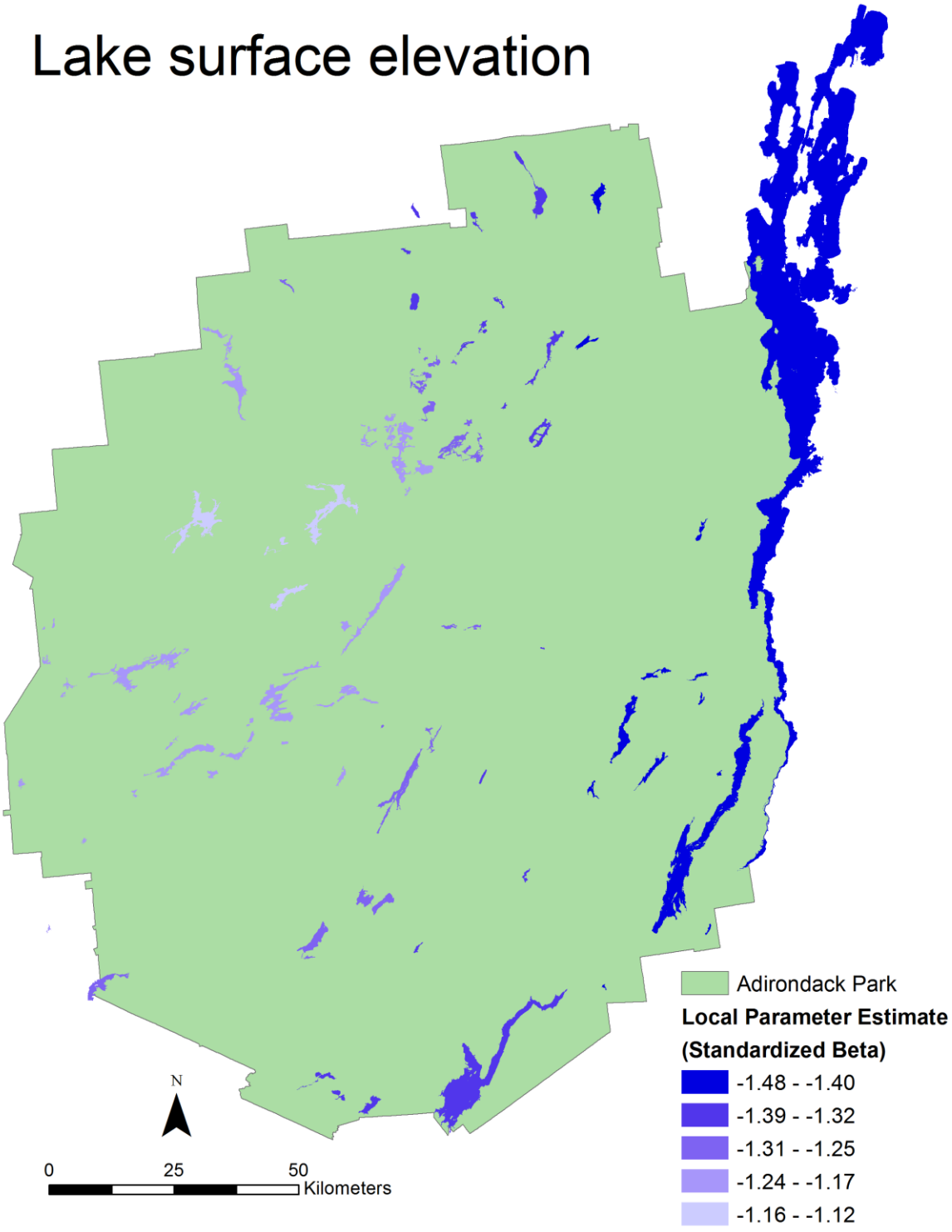


Figure 19. Parameter estimate of lake surface elevation from CLP presence-absence GWLR Model 1.

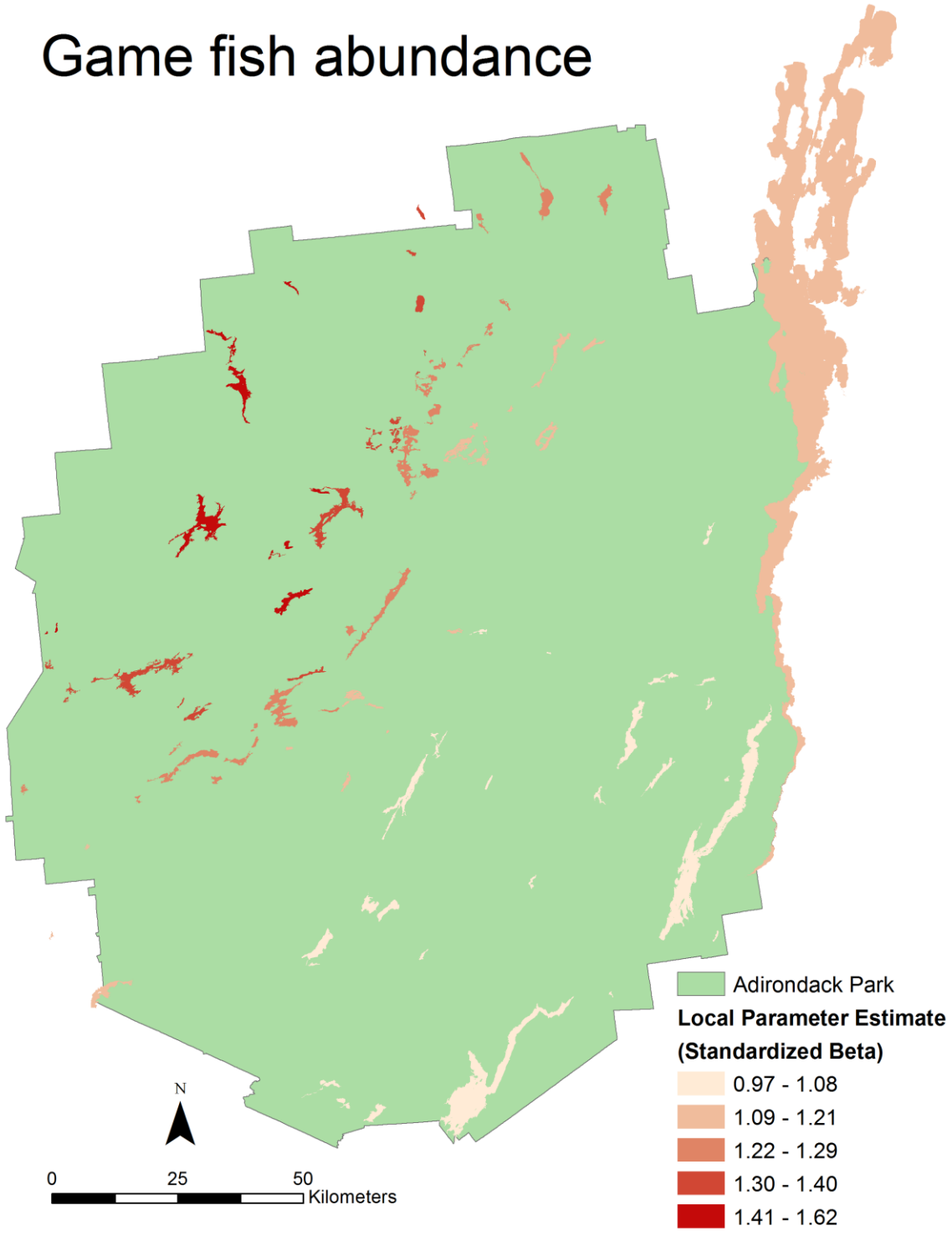


Figure 20. Parameter estimate of game fish abundance from CLP presence-absence GWLR Model 1.

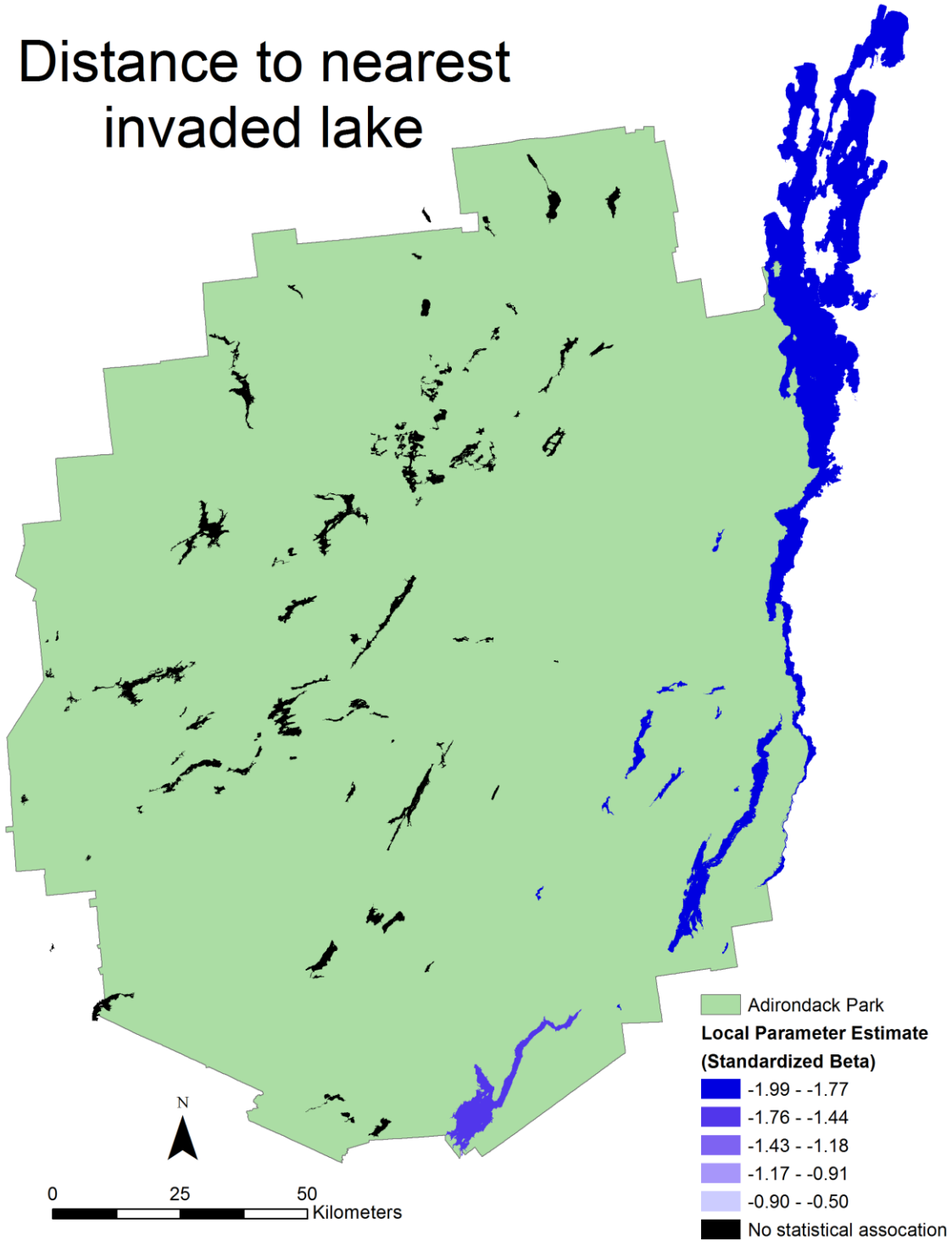


Figure 21. Parameter estimate of distance to nearest invaded lake from CLP presence-absence GWLR Model 1.

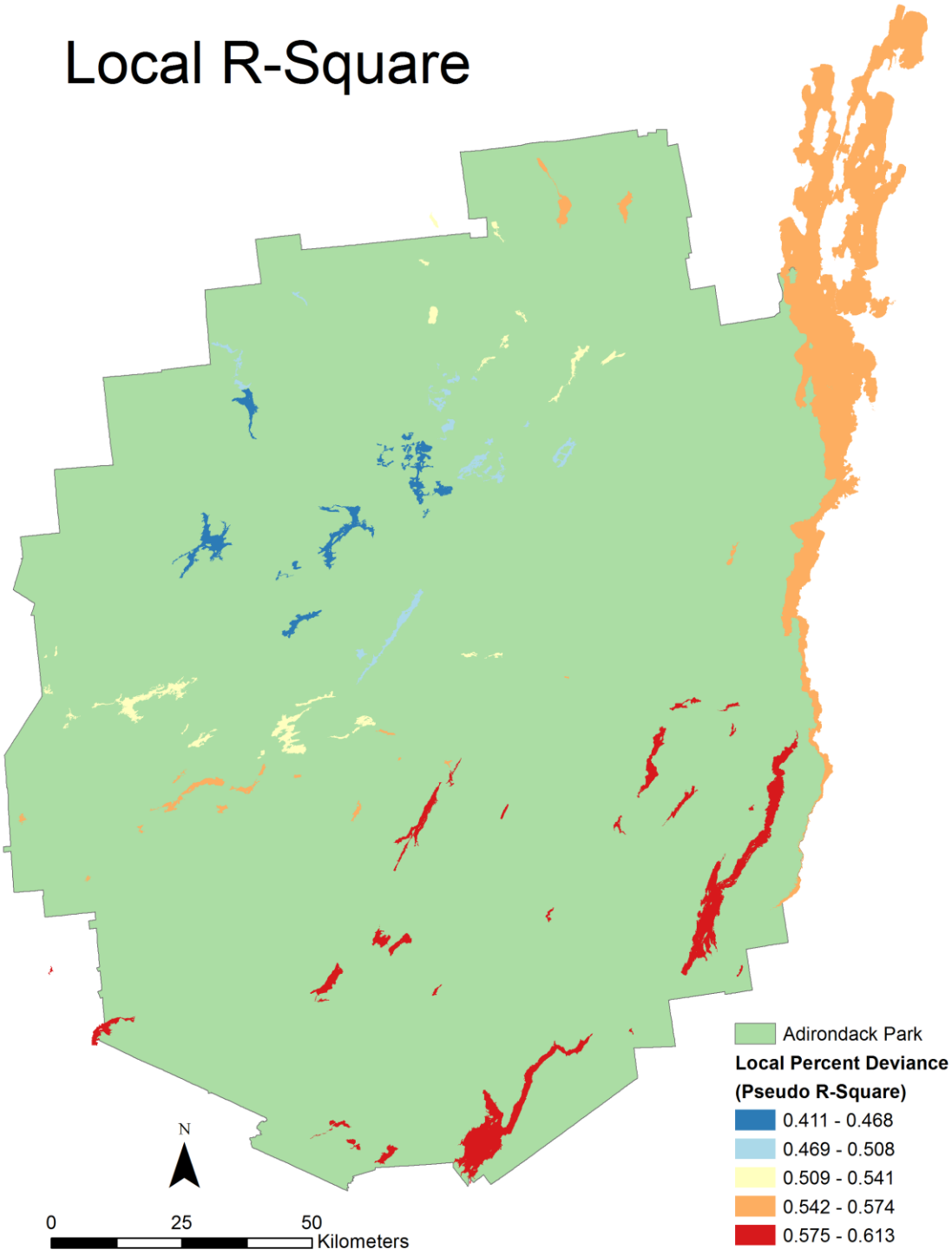


Figure 22. Local percent deviance from CLP presence-absence GWLR Model 1.

CHAPTER 5: Discussion

The following discussion relating global relationships identified by the LR are applicable in other regions as well, and findings from other studies can be compared to these. However, the GWLR findings only apply to the Adirondack Park region as it is a localized prediction.

5.1 Predicting EWM Presence-Absence

The main predictors of EWM presence-absence were found to be lake surface elevation, distance to nearest invaded lake, distance to nearest interstate highway I-87 exit, and percent developed open space. These variables were present in most models and during model selection contributed the most to reducing AICc. As indicated by standardized beta values, the higher the lake surface elevation the less likely it is that EWM would be present. This relationship may be due to colder temperatures that are present at higher elevations. Smith and Barko (1990) found that EWM reaches a maximum growth rate at 30-35 degrees Celsius and is excluded from shallow areas of lakes during times of ice cover. Ice cover may be correlative to lake surface elevation and reducing the likelihood of EWM presence, while cold air temperatures themselves are likely to not play a significant role in predicting EWM presence-absence as average temperature was not found to be a significant predictor in any models in this study and others (Nichols and Shaw, 1986). Also, EWM is a herbaceous perennial plant (root mass survives all year long) (Aiken et al., 1979). In Adirondack Park, EWM has been found to be able to maintain large biomass through New York winters in low elevation lakes like Lake George at 97 meters (Madsen et al., 1991). In the localized predictions, there was a large area in the northwest with no association to lake surface elevation and EWM presence-absence. This area is a mix of low and high elevation, and there is no clear explanation for the lack of statistical association. On the other hand, Lake Champlain and Lake George, which are in a valley, follow expected strong negative associations as seen on the map.

Furthermore, the most relevant predictor was distance to nearest invaded lake as it had the second highest R_L^2 value, was in each model, and the localized estimates were

statistically significant all across the Adirondacks. The strength of this predictor is likely associated with EWM fragmentation, commonly cited as the most important spreading mechanism of the species (Aiken et al., 1979; Madsen et al., 1991) where stems of the plant break off and float downstream (or are transported by other means) until they sink and develop roots (Dunbar, 2009; Boylen et al., 2006). Stems fragment naturally during auto fragmentation (self-generated fragmentation) which is part of EWM's life cycle and occurs after the growing season. They may also be naturally induced to fragment by wave action. Additionally, boaters may induce waves or get stems caught onto their boats or trailers and transport them to nearby lakes. Their ability to survive as fragments is well documented. Bruckerhoff et al. (2015) found that single stems of EWM can survive up to 18 hours of air exposure, and when clumped around a boat propeller can survive 2-3 days in temperate summer conditions. Laboratory experiments testing the drying of single EWM fragments have found that fragments were 40% viable after 1 hour of desiccation and 10 - 15% after 3 hours (Evans et al., 2011; Jerde et al., 2012; Barnes et al., 2013) though Evans et al. (2011) found that even when fragments were 100% dry there is still a 0.02 probability of growth and even observed root growth in fully desiccated fragments. Other studies observing fragments in water have found that 50% of the time the plant fragments are able to survive, re-sprout and produce new plants (Robinson, 2002). Madsen et al. (1988) found that under low nutrient (oligotrophic) water conditions fragments could survive and grow from 20cm to 37cm length (and 0.172 to 0.241 g dry weight) after 36 days. Therefore, EWM fragments are able to survive when exposed to air and are likely the cause of invasion to nearby lakes when transported by boats, or natural dispersion by water. It should be recognized that complementary to this predictor is open water in the riparian zone, found to be a significant predictor and indicative of freshwater connectivity. The lakes with the greatest negative association are closer in space in the northwest part of the Adirondacks and are likely well connected hydrologically.

Many predictors representing or related to human presence in this study were found to be significant predictors of the spread of EWM (Smith and Barko, 1990; Johnson et al., 2001). The strongest predictor was distance to nearest interstate highway exit I-87, found in most models and having the highest R_L^2. Lakes that are close to human communities are preferred for water tourism, and as cited previously it is known that EWM fragments

can survive when exposed to air for reasonable time periods so their transport by humans is feasible. For the Adirondack Park region, the map from the GWLR Model 1 estimate shows a large area in the south of Adirondack Park with no statistical association to nearest distance to highway I-87 exit, and the greatest negative association in the northwest region. It is not clear what is causing these patterns but one explanation may be that the northwest region is closer to the I-87 where it meets Quebec, Ontario, and Vermont bringing in tourists from these regions. Moreover, a strong association to human presence would imply that lake access should be a contributing predictor in the models which was not the case in this study as lake access was not present in any models or found to be a significant predictor during model testing, and rarely contributed to the top models found during model selection. Buchan and Padilla (2000) studied EWM presence-absence in Wisconsin lakes and similarly found lake access by public launches to be a poor predictor. However, the lake access variable may not be reflective of the spread of EWM by human means as it only records two types of lake access, carry down and public launch. A potentially more useful predictor for EWM presence-absence that should be tested in future studies is total number boat launches, both private and public, as this is fully reflective of boating accessibility. Furthermore, private boat access may be reflected in the percent developed open space in the riparian zone, which was present in most models and had the fourth highest R_L^2. This represents grass lawns built and used by humans and family homes. Since these single-unit homes are situated on the water, it is likely they have private boat access on the lake and may contribute to intralake dispersal of EWM by fragmenting or transporting the species (Buchan and Padilla, 2000). Based on the map produced by Model 1, this association is most likely to hold true in the southwest region of Adirondack Park. Human presence in the riparian zone is verified with aggregation index of developed open space, SHEI and developed medium intensity being significant predictors in a couple of models, though all having low R_L^2 values. In combination, AI of developed open space and SHEI indicate that the more aggregated human presence is in the riparian zone, or the less diverse (/even) the landscape is, then the greater the likelihood of EWM. Game fish abundance was another variable included in the analysis that was related to human presence in lakes. It was found to be a significant predictor in Model 3 however, it was not found to significantly decrease AICc during model selection.

Additional research must be undertaken to determine the magnitude of its predictive capabilities.

Mixed forest is likely another important variable, although it is not found in all the models. The model selection process was discriminatory of models with redundant variables, and thus the "top" logistic regression models from the AICc selection algorithm were not all selected. These top models all had percent mixed forest in the riparian zone as a predictor, and in the models selected for this study percent mixed forest is a significant predictor. As mixed forest increases in the riparian zone so does the likelihood of EWM presence. Interestingly, deciduous forest in the riparian zone has the opposite relationship with EWM presence, as seen in Model 2 and the bivariate regression. Shaker (2013) found the same association with mixed forest and deciduous forest in a logistic regression on EWM presence-absence. The exact reason behind such relationships is not certain and not found in the literature, however, the forest type likely influences water chemistry and nutrients in the lakes sediment, which in turn can create suitable or non-suitable habitat for EWM growth (see Madsen (1998), Smith and Barko (1990), and Nichols and Shaw (1986) for preferred life history factors). Buchan and Padilla (2000) and Cheruvelil and Soranno (2007) examined the influence of landscape features on lake macrophytes presence. The former found that forest coverage had a negative effect on EWM presence and the latter found no associations with milfoil, though did find chlorophyll *a* and agricultural lands in the riparian zone to be the greatest predictors of milfoil presence. These are mixed findings, and more research needs to be done to determine the exact influence of riparian land cover, as well as land cover at different spatial scales, on EWM presence-absence in lakes. GWLR can help discern such relationships with the ability to map local variations, such as the map for percent mixed forest showing where the associations are present.

Lastly, the only climatic factor found to be a significant predictor, albeit with an $R_L^2 = 0.03$, was mean precipitation. It has a negative association with EWM presence that is likely explained by precipitations influence on water chemistry. Tracy et al. (2003) found that during a drought in the late 1980s in Michigan rooted macrophyte communities increased because of greater light availability and associated changes in nutrient availability, which may also explain EWM's preference for less precipitation.

Additionally, precipitation may be correlated to elevation as lakes higher in elevation receive more input from precipitation and less from groundwater (Kratz et al., 1990). Lillie and Barko (1990) found EWM only in restricted groundwater input areas in Devil's Lake, Wisconsin indicating a preference for the nutrient rich waters. An interplay of these factors and others (i.e. water clarity, nutrients, etc.) were not in the scope of this work but may explain EWM's association with precipitation.

The clusters show that in general there is a significantly higher EWM presence in the north and east so these regions should be targeted as problematic areas of invasion for the nearby lakes with EWM absent. The midwest region with low-low values of local Moran's I shows the inverse to be true, and so that area may not pose as great a risk of being invaded as there is not a clustering of EWM presence.

5.2 Predicting CLP Presence-Absence

The main predictors for CLP were lake surface elevation, game fish abundance, and distance to nearest invaded lake, as all three were found in most top LR models, consistently reducing AICc when added to models, and were significant predictors in all models. Other weaker predictors that significantly contributed to a model were distance to nearest highway exit, percent open water, and RPR in the riparian zone. CLP is only present in 15 lakes relative to EWM which is present in 43 lakes which may explain the few variables that are significant in its prediction. Also, CLP propagates by seed production (turions) while EWM is more likely to disperse by fragmentation.

At higher elevations, ice and snow cover is likely a limiting factor for CLP establishment in lakes which may explain the negative association with CLP presence. Valley and Heiskary (2012) found that CLP growth was reduced in Minnesota lakes having greater amounts of snow-covered ice. Additionally, the lakes in Adirondack Park at higher elevations are more oligotrophic while at lower elevations they are mesotrophic which is preferable habitat for CLP to establish ((Boylen et al., 2006). Furthermore, the mechanisms of dispersal of CLP (of their dormant apices) are by water currents, waterfowl that eat their seeds, and entanglement on boats or their trailers (Catling and Dobson, 1985). Lakes at higher elevations are not as prone to these dispersal mechanisms relative to lakes at low elevation that may be easier to reach by humans or migrating waterfowl. For Adirondack

park, the map of lake surface elevation parameter estimates by from Model 1 of GWLR corroborates these relationships as the negative associations appear to follow elevation with the greatest association in the east which has the lowest elevations. Overall, water chemistry, decreased light availability, and inaccessibility may be a limiting factor for CLP establishment at higher elevations.

Furthermore, game fish abundance showed a positive association with CLP presence. The variable represented an absence of game fish as 0 and 3 to be three species of game fish present, and the fish counted were yellow perch, smallmouth bass, and rainbow trout. One plausible explanation for its positive association to CLP is that the macrophyte is home to many small organisms that fish feed on (Krecker, 1939) and/or provides habitat for sport fish such as northern pike, largemouth bass, and bluegill (Heiskary and Valley, 2011).

Another reason for the association may be that fishermen frequent lakes with higher abundances of game fish and accidentally introduce CLP. It should be noted that Stuckey (1979) suggested CLP may have been introduced initially with fish stocking operations, though it is not clear if this actually occurred. In any case, New York stocks waterways with sport fish (including rainbow trout, one of the game fish included in the predictor) every year (NY DEC, 2017) and since fishermen do target lakes with sport fish and their boats are likely to be a dispersal mechanism, then this presents a conservation opportunity. The waterbodies that are stocked would likely be good opportunities for field observation efforts and areas where to promote boat cleaning by APIPP. In Adirondack Park, the map of game fish abundance provided by GWLR from Model 1 would be another useful tool to identify areas where CLP dispersal is likely to occur based on the model associations.

Lastly, distance to nearest invaded lake was amongst the top predictors, corroborating the potential dispersal mechanisms by overland transport by humans (boats, trailers) or natural dispersal by water. Similar to EWM, CLP fragments can also survive when exposed to air for up to 12 hours found in one study by Bruckerhoff et al. (2015). Barnes et al. (2013) examined turion viability in a laboratory experiment and found them 55% viable after 1 hour and 10% viable after 3 hours. However, for the Adirondack Park area this relationship does not apply across the entire region. The GWLR parameter estimate map shows that most lakes did not have a statistical association with CLP

presence. The lakes with no association are in areas of higher elevation to the west of the Champlain Valley and most are in different watersheds, so the terrain is likely a barrier for CLP to naturally disperse by waterway. Lake Champlain has the strongest association with CLP presence and distance to nearest invaded lake, which is likely since it is the largest lake and most frequented by tourists. Clusters of CLP presence shown by the local Moran's I results show that the southeast and north are problematic areas as those lakes not invaded nearby are at risk of invasion.

CHAPTER 6: Conclusion

6.1 Future Work and Limitations

A key limitation of this study was not utilizing a semi-parametric model, a model incorporating both stationary and non-stationary variables. This may have further improved the model performance, as has been found and suggested by Fotheringham et al. (2002) and should be implemented in future work. Currently, the GWR4 software cannot compute semi-parametric GWLR models, though this may likely not be the case in the future. Another issue that may have occurred is overfitting when some of the models employed the use of a very small number of neighbors in the GWLR such as Model 3 for CLP presence-absence prediction which only used 16% of the total neighbors. Logistic regression is most stable when at least 10 data points are used per predictor (Babyak, 2004). Furthermore, the data set used in this study was made up of various dates collected which could have skewed the results. For instance, the land cover data is from 2011 and the response variables were from 2016. However, the Adirondack Park region is a protected area and the landscape is unlikely to have drastically changed over this period of time.

The predictors include a variety of abiotic factors, though it is recognized that there are additional abiotic and biotic variables that may enhance the quality of a predictive model. Firstly, an updated data set is always beneficial for any study. As mentioned in the discussion, if the number of private boat launches were included in the Lake Access variable then the predictor may turn out to be quite important which is unclear in this study without the inclusion of this type of access to the lake. Additionally, more hydrological

predictors may better explain the importance of precipitation and whether rainfall, snow, or ice cover play a role in EWM and CLP prediction. Groundwater input and evaporation rates in the lake may provide other clues as to what makes a lake suitable habitat specifically for EWM presence-absence as CLP did not have any associations to a climatic factor. For the prediction of CLP presence, the inclusion of a predictor representing waterfowl presence at the lakes would be the most important predictor to include as studies have indicated the ability of migrating waterfowl to act as a secondary dispersal mechanism (Green, 2016). Other predictors that could be included are trophic index (or variables that are representative of trophic index like pH and total phosphporus), sediment rates, and turbidity. EWM has been found to prefer lakes of a certain trophic index with a lower boundary of 36 and upper of 74 for Carlson's trophic index (Madsen, 1998). Also for EWM, sediment deposition and turbidity following storm events (Mataraza et al. 1999), and the presence of predators like *Acentria ephemeralla* which feeds on EWM (Johnson et al., 2000) have an influence on macrophyte populations. Obtaining such data was out of scope for this work. However, as noted in Shaker et al. (2017), this may not be a significant limitation of the study because of the collinearity found in other studies between biological, chemical, and geological lake parameters and surrounding landscape characteristics. In an ideal model, all these predictors would be considered.

6.2 Summary

In this study, the aim was to explore a variety of lake and riparian zone landscape factors, and climate variables to assess whether there existed a significant predictive relationship between EWM and CLP presence-absence which would shed light on potential future colonization sites of these species in order to help management plans of IS prevention and containment. A secondary aim was to test whether the predictions were better forecast by LR or GWLR, where the former did not take into account spatial variation of the predictors while the latter did. Some of the key predictors expected to be significant were found to be significant, such as those indicating human presence and natural connectivity of waterways, while others were surprisingly insignificant such as the limited association of AIS presence-absence with climactic factors (e.g. temperature) and lake morphology. There existed a large suite of significant predictors for EWM which verified

that EWM is more widespread and an aggressive aquatic invader, making it complex to accurately predict. In contrast, CLP is less widespread and less present in Adirondack Park with fewer predictors.

Overall, the GWLR modeling technique showed improvement in model diagnostics when compared to LR. The results also reiterate those found in Shaker et al. (2017) whom found GWR to improve model performance when predicting Aquatic Invasive Species Richness in Adirondack Park using the same sets of predictors. For EWM presence-absence this proved to be truer than for CLP presence absence: explained variation of the response variable in the EWM models increased by 23% when using GWLR, and only 16% for CLP models. GWLR should certainly be used as an exploratory modeling technique to investigate if it improves model performance, and especially if spatial autocollinearity is a problem in the residuals of a model or is found amongst the variables. In this study, spatial autocollinearity was an issue with Model 2 for EWM in the LR model and the GWLR was able to deal with the problem. Overall, it should be expected that geographic weighted regression would improve model performance compared to a global regression model.

The assumptions of logistic regression were met in this study, and confirmed by many diagnostics. The key advantage of GWLR was its naturally-occurring spatially-varying predictive capability, enabling an intuitive platform from which to interpret the results of the model. Rather than rely on a single value from logistic regression indicating the "overall" impact of a certain predictor, GWLR enables deeper understanding of the association of the predictor with the local geography – a helpful property for highlighting certain problematic areas *within* a region. In the case of the APIPP, for example, this additional "wrinkle" in the predictive capability can be useful for highlighting which areas may be in the most precarious situations and may serve as an initial point-of-contact for implementing pilot programs for combatting the spread of EWM and CLP. This can be done without having to immediately spend those resources across the entire region, or without implementing these programs in areas where they may be less beneficial, key strengths of the GWLR approach.

The use of open data, software, and citizen science data makes this study unique. Using open access data and software to be able to predict invasion of AIS for this study came at no cost, which implies that there may be little cost needed for predicting AIS invasions or similar work. This is extremely important as AIS management by application of herbicides and clean ups already costs millions of dollars yearly (Pimental et al., 2005). The APIPP data is updated yearly on presence-absence, and most of the other variables came from governmental agencies that update as frequently as possible. There is always the concern that volunteers may incorrectly identify species and Crall et al. (2011) did find that volunteers could identify higher taxonomic groups but not different species as well as scientists could. There needs to be good training programs and collaboration with scientists to obtain the best data possible. Cacho et al. (2010) found through a simulated spatial model of a hypothetical invasion that the combination of passive surveillance, when the public reports to higher authorities of IS sightings, and active surveillance (actively searching for IS) to be the only way to fully stop invasive species spread because of the speed at which IS are identified and the funds required to combat the IS. Furthermore, low cost spatial analysis can be part of the solution to IS as it can speed up active surveillance. Once data has been collected, it needs to be managed and stored properly, analyzed and reported and so "to assist decision makers with complex spatial problems, geoprocessing systems must support decision research process by providing the decision maker with a flexible, problem solving environment" (p. 403 Densham, 1991). The framework for this is provided by spatial decision support systems (SDSS) (Densham, 1991). This study shows that the SDSS framework may not require a lot of money to be implemented. In fact, the applications used in the study were mostly free including GWR4 software, SAM ecology software, R Studio, and QGIS. The only software that was not free of cost was ArcGIS but that software is available free of charge for most students and other GIS technologies can be used in its place.

Appendices

Appendix A – Data Source

Variable	Units/Information	Source	Classification
Presence-absence of species	0 = absent 1 = present	APIPP	ordinal
Lake area (A)	sq km	NYDEC	continuous
Lake perimeter (P)	km	NYDEC	continuous
P-A Ratio	km/sq km	NYDEC	continuous
Maximum depth	m	SCFG	continuous
Lake surface elevation	m	USGS	continuous
Lake access type	1) carry down only 2) public launch	SCFG	ordinal
Distance to nearest invaded lake	km	APIPP	continuous
Game fish abundance: yellow perch, smallmouth bass, rainbow trout	0) absent 1) one species 2) two species 3) three species	SCFG	ordinal
Distance to I-87 exit	km	DOT	continuous
Distance to nearest populated place	km	CENSUS	continuous
Climate			
Temperature (mean, range)	ºC	USDA - NRCS	continuous
Precipitation (mean, range)	inches	USDA - NRCS	continuous
Land cover composition ϕ	percentage of total	USGS	continuous
Developed, open space (DO)	%		
Developed, low intensity	%		
Developed, medium intensity	%		
Deciduous forest	%		
Evergreen forest (EF)	%		
Mixed forest	%		
Pasture/hay	%		
Cultivated crops	%		
Woody wetlands	%		
Emergent herbaceous wetlands	%		
Land cover class configuration ϕ		USGS	continuous
AI, DO	%		
AI, EF	%		
PLADJ, DO	%		
PLADJ, EF	%		
AREA_AM, DO	sq m		
AREA_AM, EF	sq m		
ENN_AM, DO	m		
ENN_AM, EF	m		
Landscape diversity ϕ		USGS	continuous
RPR	%		
SHDI	SHDI ≥ 0, w/o limit		
SHEI	0 ≤ SHEI ≤ 1		

ϕ Landscapes calculated using a 300 meter buffer for each lake.

Notes: APIPP = Adirondack Park Invasive Plant Program, CENSUS = U.S. Census, DOT = Department of Transportation, NYDEC = New York State Department of Environmental Conservation, SCFG = Sportsman's Connection Fishing Guide, USDA – NRC = United States Department of Agriculture – Natural Resources Conservation Service, USGS = U.S. Geological Survey, AI = Aggregation Index, PLADJ = Percentage of Like Adjacencies, AREA_AM = Area-Weighted Mean Patch Area, ENN_AM = Area-Weighted Mean Euclidean Nearest Neighbor Distance, RPR = Relative Patch Richness, SHDI = Shannon's Diversity Index, SHEI = Shannon's Evenness Index.

National land cover data description from 2011. (USGS, 2014).

Open Water- areas of open water, generally with less than 25% cover of vegetation or soil.
Perennial Ice/Snow- areas characterized by a perennial cover of ice and/or snow, generally greater than 25% of total cover.
Developed, Open Space- areas with a mixture of some constructed materials, but mostly vegetation in the form of lawn grasses. Impervious surfaces account for less than 20% of total cover. These areas most commonly include large-lot single-family housing units, parks, golf courses, and vegetation planted in developed settings for recreation, erosion control, or aesthetic purposes.
Developed, Low Intensity- areas with a mixture of constructed materials and vegetation. Impervious surfaces account for 20% to 49% percent of total cover. These areas most commonly include single-family housing units.
Developed, Medium Intensity -areas with a mixture of constructed materials and vegetation. Impervious surfaces account for 50% to 79% of the total cover. These areas most commonly include single-family housing units.
Developed High Intensity-highly developed areas where people reside or work in high numbers. Examples include apartment complexes, row houses and commercial/industrial. Impervious surfaces account for 80% to 100% of the total cover.
Barren Land (Rock/Sand/Clay) - areas of bedrock, desert pavement, scarps, talus, slides, volcanic material, glacial debris, sand dunes, strip mines, gravel pits and other accumulations of earthen material. Generally, vegetation accounts for less than 15% of total cover.
Deciduous Forest- areas dominated by trees generally greater than 5 meters tall, and greater than 20% of total vegetation cover. More than 75% of the tree species shed foliage simultaneously in response to seasonal change.
Evergreen Forest- areas dominated by trees generally greater than 5 meters tall, and greater than 20% of total vegetation cover. More than 75% of the tree species maintain their leaves all year. Canopy is never without green foliage.
Mixed Forest- areas dominated by trees generally greater than 5 meters tall, and greater than 20% of total vegetation cover. Neither deciduous nor evergreen species are greater than 75% of total tree cover.
Dwarf Scrub- Alaska only areas dominated by shrubs less than 20 centimeters tall with shrub canopy typically greater than 20% of total vegetation. This type is often co-associated with grasses, sedges, herbs, and non-vascular vegetation.
Shrub/Scrub- areas dominated by shrubs; less than 5 meters tall with shrub canopy typically greater than 20% of total vegetation. This class includes true shrubs, young trees in an early successional stage or trees stunted from environmental conditions.
Grassland/Herbaceous- areas dominated by gramanoid or herbaceous vegetation, generally greater than 80% of total vegetation. These areas are not subject to intensive management such as tilling, but can be utilized for grazing.
Pasture/Hay-areas of grasses, legumes, or grass-legume mixtures planted for livestock grazing or the production of seed or hay crops, typically on a perennial cycle. Pasture/hay vegetation accounts for greater than 20% of total vegetation.
Cultivated Crops -areas used for the production of annual crops, such as corn, soybeans, vegetables, tobacco, and cotton, and also perennial woody crops such as orchards and vineyards. Crop vegetation accounts for greater than 20% of total vegetation. This class also includes all land being actively tilled.
Woody Wetlands- areas where forest or shrubland vegetation accounts for greater than 20% of vegetative cover and the soil or substrate is periodically saturated with or covered with water.
Emergent Herbaceous Wetlands- Areas where perennial herbaceous vegetation accounts for greater than 80% of vegetative cover and the soil or substrate is periodically saturated with or covered with water.

Appendix B – Distribution of Data

Skewness and kurtosis (calculated in *SAM* ecology software*)* of continuous predictor data post and prior transformations (*N* = 126). -- No transformation applied.

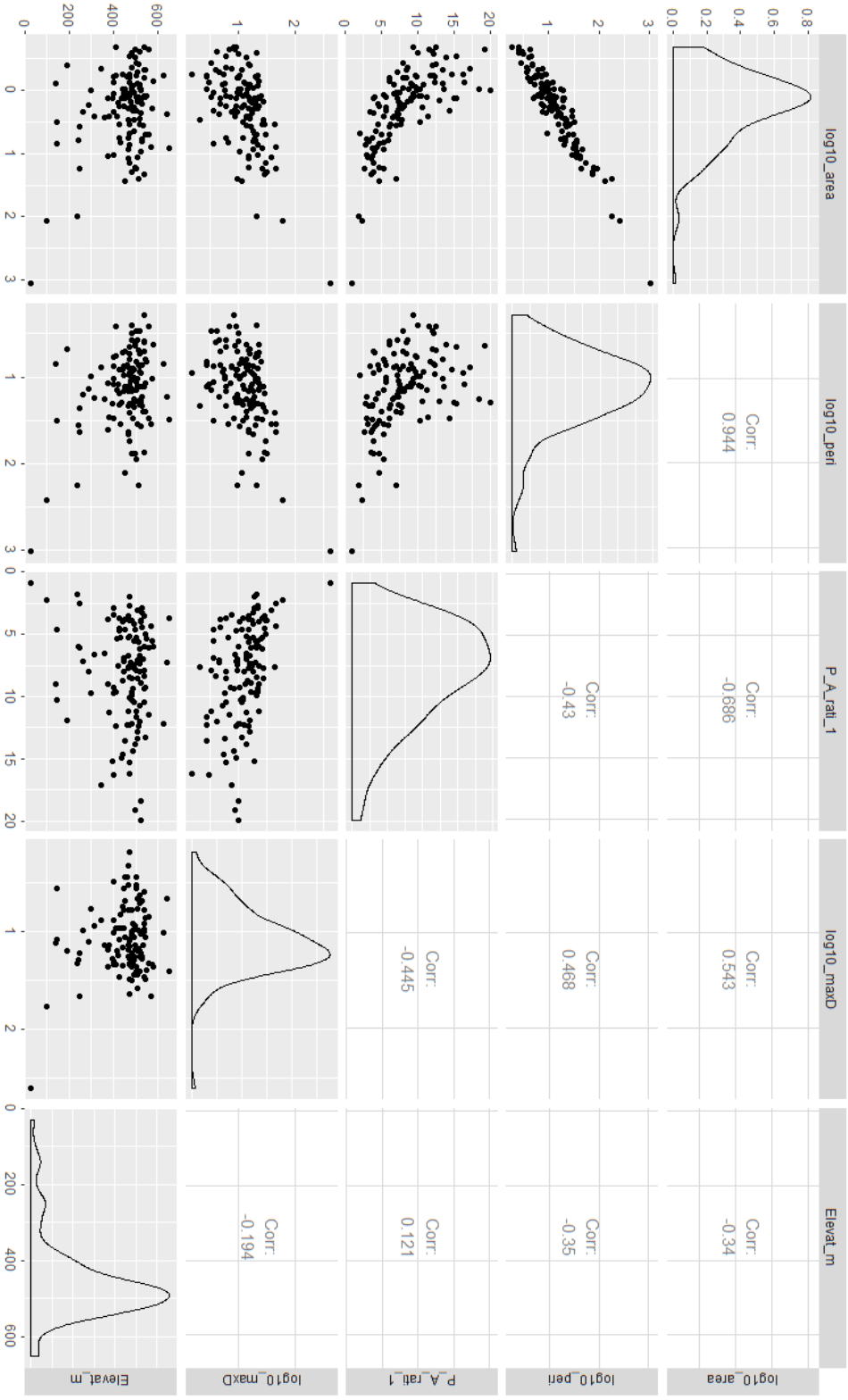
Predictor Variable	Units	Skewness	Kurtosis	Skewness Post-Transform	Kurtosis Post-Transform
Surface area [a]	sq. km	10.911	121.034	1.239	3.125
Perimeter [a]	km	9.300	95.242	1.045	2.435
Perimeter-area ratio	km/sq. km	0.715	0.131	---	---
Maximum depth [a]	m	9.997	107.39	0.262	2.608
Surface elevation	m	-1.722	3.537	---	---
Distance to nearest invaded lake	km				
Eurasian watermilfoil		0.662	-0.694	---	---
Curly-leaf pondweed		0.717	0.391	---	---
Distance to I-87 highway exit	km	0.167	-0.169	---	---
Distance to nearest populated place	km	1.322	3.119	---	---
Average Temperature [a]		1.983	4.041	1.656	2.794
Max Average Temp [c]		2.239	6.288	-1.486	3.726
Min Average Temp		1.404	1.733	---	---
Temperature Range		-0.582	0.373	---	---
Average Precipitation		0.382	0.44	---	---
Max Average Precipitation		0.752	0.45	---	---
Min Average Precipitation		0.022	1.068	---	---
Range Precipitation [b]		4.599	26.821	1.78	4.264
Land Cover [ϕ]	Percent of total area				
Developed, open space [b] (DO)	%	2.329	6.693	0.612	0.226
Developed, low intensity [b]	%	3.663	17.022	1.450	1.976
Developed, medium intensity [b]	%	4.781	28.563	2.023	4.724
Developed, high intensity [b, ×]	%	4.537	22.056	3.183	9.815
Deciduous forest	%	0.522	-0.135	---	---
Evergreen forest (EF)	%	0.443	-0.42	---	---
Mixed forest [b]	%	2.719	9.566	1.012	2.237
Pasture and hay [b, ×]	%	6.053	40.686	3.723	15.13
Cultivated crops [b, ×]	%	5.036	26.192	3.680	13.693
Shrub and scrubland [b]	%	3.519	15.67	1.282	2.587
Herbaceous [b, ×]	%	4.24	19.641	2.282	5.34
Emergent herbaceous wetland [b]	%	5.111	30.006	2.188	6.584
Woody wetland [b]	%	2.633	8.752	1.042	1.518
Open water	%	0.624	0.354	---	---
Barren [b, ×]	%	6.097	41.552	4.551	21.143
Land cover class metric [ϕ]					
AI, DO	%	-1.126	0.102	---	---
AI, EF	%	-0.610	0.727	---	---
PLADJ, DO	%	-0.950	-0.262	---	---
PLADJ, EF	%	-0.858	1.174	---	---
AREA_AM, DO [b]	sq. m	3.099	13.792	0.908	1.118

AREA_AM, EF [b]	sq. m	1.865	3.866	-0.255	-0.219
ENN_AM, DO [d]	m	10.353	112.484	-0.957	-0.323
ENN_AM, EF [a]	m	2.663	9.117	0.385	3.308
Landscape diversity [ϕ]					
RPR	%	0.271	-0.301	---	---
SHDI	SHDI ≥ 0, w/o limit	0.480	0.936	---	---
SHEI	0 ≤ SHEI ≤ 1	-0.149	0.360	---	---

[ϕ] Calculated within a 300 meter buffer of the lakes.
[a] Variable transformed by $\log_{10}$ (variable).
[b] Variable transformed by square-root.
[c] Variable transformed by 1/ (variable)^2.
[d] Variable transformed by $\log_{10}$ (variable + 1).
[×] Variable is not used in follow-up statistical procedures because coverage is minimal (0.5% or less) and/or the variable is not normally distributed even after transformations.

Appendix C – Multicollinearity Diagnostics

C.1 Pearson Correlations and Scatterplots



LogMeanTempC
MaxTempC_ReciprocalSqr
Min_TempC
Range_TempC
MEAN_PPT
MAX_PPT
MIN_PPT
RangePptSqrt
Corr:
-0.918
Corr:
0.96
Corr:
-0.29
Corr:
-0.16
Corr:
-0.0535
Corr:
-0.216
Corr:
0.331
Corr:
-0.77
Corr:
-0.107
Corr:
0.256
Corr:
0.143
Corr:
0.313
Corr:
-0.353
Corr:
-0.541
Corr:
-0.062
Corr:
0.0274
Corr:
-0.112
Corr:
0.277
Corr:
-0.268
Corr:
-0.234
Corr:
-0.275
Corr:
0.0797
Corr:
0.956
Corr:
0.984
Corr:
-0.0545
Corr:
0.9
Corr:
0.194
Corr:
-0.209

C.2 Spearman Correlations

	lper	parat	lmdep	elev	acc	gam	ewid2	clpd	di87	dpop	mtemp	rant
lper												
parat	-0.46****											
lmdep	0.38****	-0.45****										
elev	-0.19*	0.05	0.02									
acc	0.44****	-0.49****	0.37****	-0.15								
gam	0.46****	-0.48****	0.42****	-0.07	0.48****							
ewid2	0.19*	0.03	-0.04	-0.06	0.04	-0.12						
clpd	0.00	0.02	0.07	0.30****	-0.01	-0.07	0.41****					
di87	-0.04	0.15	0.01	0.24**	-0.08	-0.19*	0.15	0.53****				
dpop	0.20*	-0.23*	0.08	-0.06	0.18*	0.00	-0.02	-0.09	0.04			
mtemp	0.07	-0.13	0.06	-0.16	0.15	0.03	-0.20*	-0.23*	-0.17	-0.03		
rant	0.27**	-0.16	0.09	-0.05	0.09	0.10	0.10	0.01	-0.09	0.07	-0.42****	
mppt	-0.13	0.14	-0.09	0.13	-0.07	-0.08	-0.06	0.23**	0.23**	0.00	-0.18*	-0.19*
ranp	0.10	-0.14	0.03	-0.06	0.03	0.06	-0.18*	-0.22*	-0.06	0.00	0.19*	0.03
devo	0.13	-0.14	0.08	-0.10	0.30****	0.29**	-0.10	-0.06	-0.39****	-0.32****	0.17	0.06
devl	0.26**	-0.10	0.02	-0.27**	0.22*	0.28**	0.07	0.03	-0.30****	-0.18*	0.14	0.07
devm	0.30****	-0.19*	0.07	-0.24**	0.34****	0.33****	0.00	-0.01	-0.21*	-0.15	0.13	0.10
decf	-0.17	0.10	0.05	0.30****	-0.08	-0.17	0.19*	0.39****	0.53****	0.14	-0.20*	0.08
evef	-0.34****	0.40****	-0.25**	-0.07	-0.18*	-0.21*	-0.13	-0.44****	-0.35****	-0.18*	0.12	-0.20*
mixf	-0.12	0.15	-0.16	0.06	-0.20*	-0.19*	-0.27**	0.10	-0.10	0.03	0.05	-0.08
sscr	0.13	0.23*	-0.05	0.27**	-0.04	0.03	0.08	0.42****	0.43****	-0.21*	-0.11	0.02
eherb	0.29****	-0.04	-0.12	-0.07	0.04	0.04	0.11	0.03	0.06	-0.13	-0.08	0.02
woodw	-0.01	0.13	-0.34****	-0.15	-0.06	-0.08	0.15	0.23*	0.06	-0.09	-0.09	0.01
open	0.60****	-0.82****	0.49****	0.01	0.45****	0.50****	-0.09	-0.07	-0.09	0.22*	0.16	0.14
aido	0.23*	-0.32****	0.22*	0.02	0.24**	0.24**	-0.13	-0.04	-0.27**	-0.11	0.18*	0.02
aief	-0.12	0.11	-0.10	0.06	-0.02	-0.09	-0.14	-0.35****	-0.25**	-0.05	0.08	-0.08
plado	0.32****	-0.40****	0.28**	-0.03	0.35****	0.35****	-0.10	-0.01	-0.30****	-0.13	0.19*	0.07
plaef	0.01	0.06	-0.07	0.00	0.04	-0.03	-0.11	-0.38****	-0.30****	-0.03	0.09	-0.05
ardo	0.40****	-0.41****	0.32****	-0.05	0.40****	0.44****	-0.05	0.01	-0.32****	-0.17	0.18*	0.11
aref	0.17	0.03	0.02	-0.17	0.01	0.06	-0.05	-0.45****	-0.39****	0.07	0.12	-0.03
endo	0.26**	-0.36****	0.15	-0.04	0.33****	0.27**	-0.01	-0.01	-0.06	0.04	0.17	0.10
enef	0.15	-0.15	0.18*	0.20*	0.02	-0.05	0.08	0.41****	0.38****	0.12	-0.15	0.14
rpr	0.62****	-0.37****	0.18*	-0.26**	0.46****	0.43****	0.16	0.02	-0.15	-0.11	0.11	0.17
shdi	-0.14	0.38****	-0.37****	-0.06	-0.05	-0.12	0.01	0.18*	-0.01	-0.37****	-0.09	0.03
shei	-0.61****	0.60****	-0.47****	0.08	-0.37****	-0.42****	-0.08	0.13	0.03	-0.25**	-0.19*	-0.09

*significant at the 90% level
**significant at the 95% level
**** significant at the 99% level

	mppt	ranp	devo	devl	devm	decf	evef	mixf	sscr	eherb	woodw	open
lper												
parat												
lmdep												
elev												
acc												
gam												
ewid2												
clpd												
di87												
dpop												
mtemp												
rant												
mppt												
ranp	0.00											
devo	-0.05	0.05										
devl	-0.01	-0.01	0.66****									
devm	-0.07	0.13	0.60****	0.74****								
decf	0.12	-0.03	-0.28**	-0.26**	-0.25**							
evef	-0.11	-0.05	0.02	-0.10	-0.14	-0.55****						
mixf	0.12	0.13	-0.22*	-0.07	-0.13	-0.12	0.09					
sscr	0.03	-0.05	-0.01	0.05	0.03	0.12	-0.23**	0.06				
eherb	0.05	0.02	0.07	0.17	0.16	-0.15	-0.10	-0.18*	0.19*			
woodw	0.17	-0.09	0.03	0.12	0.09	-0.05	-0.12	-0.03	0.05	0.38****		
open	-0.10	0.08	0.12	0.10	0.18*	-0.19*	-0.38****	-0.20*	-0.05	0.06	-0.23**	
aido	-0.12	0.05	0.64****	0.43****	0.38****	-0.23**	-0.09	-0.10	-0.07	0.03	-0.17	0.34****
aief	-0.09	0.03	0.01	-0.09	-0.10	-0.49****	0.73****	-0.05	-0.08	0.06	-0.16	-0.09
plado	-0.10	0.06	0.75****	0.53****	0.50****	-0.26**	-0.13	-0.15	-0.09	0.06	-0.13	0.39****
plaef	-0.14	0.05	0.03	-0.04	-0.05	-0.56****	0.74****	-0.07	-0.10	0.09	-0.16	-0.02
ardo	-0.08	0.10	0.82****	0.59****	0.58****	-0.29**	-0.12	-0.20*	-0.09	0.06	-0.11	0.40****
aref	-0.17	-0.03	0.01	0.03	-0.05	-0.67****	0.72****	-0.02	-0.24**	0.00	-0.17	0.06
endo	-0.13	-0.03	0.32****	0.14	0.10	-0.02	-0.13	-0.14	-0.08	0.00	-0.04	0.34****
enef	0.19*	-0.06	-0.05	-0.05	0.01	0.50****	-0.70****	-0.04	0.25**	0.06	-0.03	0.15
rpr	-0.09	0.10	0.57****	0.73****	0.70****	-0.23**	-0.26**	-0.22*	0.15	0.40****	0.14	0.41****
shdi	0.01	0.03	0.34****	0.33****	0.29****	-0.09	0.14	0.32****	0.38****	0.23**	0.39****	-0.45****
shei	0.13	-0.06	-0.12	-0.19*	-0.19*	0.06	0.32****	0.46****	0.12	-0.07	0.22*	-0.73****

*significant at the 90% level
**significant at the 95% level
**** significant at the 99% level

	aido	aief	plado	plaef	ardo	aref	endo	enef	rpr	shdi
lper										
parat										
lmdep										
elev										
acc										
gam										
ewid2										
clpd										
di87										
dpop										
mtemp										
rant										
mppt										
ranp										
devo										
devl										
devm										
decf										
evef										
mixf										
sscr										
eherb										
woodw										
open										
aido										
aief	0.04									
plado	0.95****	-0.02								
plaef	0.07	0.98****	0.03							
ardo	0.81****	-0.06	0.92****	0.00						
aref	0.09	0.79****	0.06	0.85****	0.05					
endo	0.32****	-0.01	0.37****	0.02	0.38****	0.03				
enef	0.00	-0.42****	0.00	-0.45****	-0.01	-0.57****	0.03			
rpr	0.42****	-0.11	0.54****	-0.02	0.62****	0.05	0.31****	0.05		
shdi	0.07	0.06	0.08	0.03	0.08	-0.06	-0.06	-0.07	0.24**	
shei	-0.25**	0.12	-0.32****	0.04	-0.37****	-0.08	-0.35****	-0.06	-0.50****	0.64****

*significant at the 90% level
**significant at the 95% level
**** significant at the 99% level

C.3 Correlation Codes from R Statistical Software

C.3.1 Pearson Correlation Code for Graph Plots

#retrieved http://www.sthda.com/english/wiki/ggally-r-package-extension-to-ggplot2-for-correlation-matrix-and-survival-plots-r-software-and-data-visualization

#Ggpairs code
ggpairs(data, columns = 1:ncol(data), title = "", axisLabels = "show", columnLabels = colnames(data[, columns]))

#example:
ggpairs(LakeTraits, diag=list(continuous="density", discrete="bar"), axisLabels="show")

C.3.2 Spearman Correlation Code for Latex Table

#Retrieved from http://www.sthda.com/english/wiki/elegant-correlation-table-using-xtable-r-package

```
corstars <-function(x, method=c("spearman"), removeTriangle=c("upper", "lower"), result=c ("latex")){

        #Compute correlation matrix
        require(Hmisc)
        x <- as.matrix(x)
        correlation_matrix<-rcorr(x, type=method[1])
        R <- correlation_matrix$r # Matrix of correlation coeficients
        p <- correlation_matrix$P # Matrix of p-value

        ## Define notions for significance levels; spacing is important. mystars <- ifelse(p < .0001,
        "****", ifelse(p < .001, "*** ", ifelse(p < .01, "** ", ifelse(p < .05, "* ", " ")))

        ## trunctuate the correlation matrix to two decimal
        R <- format(round(cbind(rep(-1.11, ncol(x)), R), 2))[,-1]

        ## build a new matrix that includes the correlations with their apropriate stars Rnew <-
        matrix(paste(R, mystars, sep=""), ncol=ncol(x)) diag(Rnew) <- paste(diag(R), " ", sep="")
        rownames(Rnew) <- colnames(x) colnames(Rnew) <- paste(colnames(x), "", sep="")

        ## remove upper triangle of correlation matrix if(removeTriangle[1]=="upper"){ Rnew <-
        as.matrix(Rnew) Rnew[upper.tri(Rnew, diag = TRUE)] <- "" Rnew <- as.data.frame(Rnew) }

        ## remove lower triangle of correlation matrix else if(removeTriangle[1]=="lower"){ Rnew <-
        as.matrix(Rnew) Rnew[lower.tri(Rnew, diag = TRUE)] <- "" Rnew <- as.data.frame(Rnew) }

        ## remove last column and return the correlation matrix
        Rnew <- cbind(Rnew[1:length(Rnew)-1])
        if (result[1]=="none") return(Rnew)
        else{
                if(result[1]=="html") print(xtable(Rnew), type="html")
                else print(xtable(Rnew), type="latex")
                }
}
```

C.4 Variance Inflation Factor

C.4.1 Variance Inflation Factor Code in R

```
#Stepwise selection VIF function returned list of variables to be kept in model
#code retrieved from https://www.r-bloggers.com/collinearity-and-stepwise-vif-selection/

> vif_func<-function(in_frame,thresh=10,trace=T,...){
+
+    require(fmsb)
+
+    if(class(in_frame) != 'data.frame') in_frame<-data.frame(in_frame)
+
+    #get initial vif value for all comparisons of variables
+    vif_init<-NULL
+    var_names <- names(in_frame)
+    for(val in var_names){
+        regressors <- var_names[-which(var_names == val)]
+        form <- paste(regressors, collapse = '+')
+        form_in <- formula(paste(val, '~', form))
+        vif_init<-rbind(vif_init, c(val, VIF(lm(form_in, data = in_frame, ...))))
+    }
+    vif_max<-max(as.numeric(vif_init[,2]), na.rm = TRUE)
+
+    if(vif_max < thresh){
+        if(trace==T){ #print output of each iteration
+            prmatrix(vif_init,collab=c('var','vif'),rowlab=rep('',nrow(vif_init)),quote=F)
+            cat('\n')
+            cat(paste('All variables have VIF < ', thresh,', max VIF ',round(vif_max,2), sep=''),'\n\n')
+        }
+        return(var_names)
+    }
+    else{
+
+        in_dat<-in_frame
+
+        #backwards selection of explanatory variables, stops when all VIF values are below 'thresh'
+        while(vif_max >= thresh){
+
+            vif_vals<-NULL
+            var_names <- names(in_dat)
+
+            for(val in var_names){
+                regressors <- var_names[-which(var_names == val)]
+                form <- paste(regressors, collapse = '+')
+                form_in <- formula(paste(val, '~', form))
+                vif_add<-VIF(lm(form_in, data = in_dat, ...))
+                vif_vals<-rbind(vif_vals,c(val,vif_add))
+            }
+            max_row<-which(vif_vals[,2] == max(as.numeric(vif_vals[,2]), na.rm = TRUE))[1]
+
+            vif_max<-as.numeric(vif_vals[max_row,2])
+
+            if(vif_max<thresh) break
+
+            if(trace==T){ #print output of each iteration
```

```
+                prmatrix(vif_vals,collab=c('var','vif'),rowlab=rep('',nrow(vif_vals)),quote=F)
+                cat('\n')
+                cat('removed: ',vif_vals[max_row,1],vif_max,'\n\n')
+                flush.console()
+            }
+
+            in_dat<-in_dat[,!names(in_dat) %in% vif_vals[max_row,1]]
+
+        }
+
+        return(names(in_dat))
+
+    }
+
+ }

#the result with the transformed data (N=126)
> vif_func(in_frame=transformedDataShortCSV, thresh=10,trace=T)
```

Appendix D – Global Moran's I Spatial Autocorrelation

Variable	Units	Moran's Index	z-score	p-value
Predictor Variables				
Lake Morphology				
Lake area	sq. km			
Perimeter	km	0.034	1.317	0.188
Perimeter-area ratio	km/sq. km			
Maximum depth	m	0.046	1.709	0.088
Surface elevation	m	0.242	7.867	0.000
Other Lake Traits				
Access type	1) carry down only 2) public launch	---	---	---
Distance to invaded lake				
Eurasian watermilfoil	km	0.425	13.449	0.000
Curly-leaf pondweed	km	0.673	21.206	0.000
Game fish abundance: yellow perch, smallmouth bass, rainbow trout	0) absent 1) one species 2) two species 3) three species	---	---	---
Distance to I-87 exit	km	0.336	10.695	0.000
Distance to nearest populated place	km	0.0698	2.449	0.0143
Climate				
Average temperature	°C	0.0199	0.878	0.379
Maximum temperature	°C			
Minimum temperature	°C			
Range temperature	°C	-0.0514	-1.353	0.176
Average precipitation	inches	0.0147	0.708	0.479
Maximum precipitation	inches			
Minimum precipitation	inches			
Range precipitation	inches	-0.0248	-0.532	0.595
Land Cover ϕ	Percent of total area			
Developed, open space (DO)	%	0.103	3.457	0.001
Developed, low intensity	%			
Developed, medium intensity	%	0.0174	0.802	0.422

Developed, high intensity	%			
Deciduous forest	%	0.212	6.836	0.000
Evergreen forest (EF)	%			
Mixed forest	%	0.204	6.635	0.000
Pasture and hay	%			
Cultivated crops	%			
Shrub and scrubland	%	0.224	7.2813	0.000
Herbaceous	%			
Emergent herbaceous wetland	%	-0.003	0.148	0.88
Woody wetland	%	0.244	7.883	0.000
Open water	%	0.008	0.525	0.599
Barren	%			
Land cover class metric ϕ				
AI, DO	%	0.118	3.931	0.000
AI, EF	%	0.199	6.454	0.000
PLADJ, DO	%			
PLADJ, EF	%			
AREA_AM, DO	sq. m			
AREA_AM, EF	sq. m			
ENN_AM, DO	m	0.088	2.978	0.003
ENN_AM, EF	m	0.218	7.129	0.000
Landscape diversity ϕ				
RPR	%	0.0809	2.762	0.006
SHDI	SHDI ≥ 0, w/o limit			
SHEI	0 ≤ SHEI ≤ 1	0.026	1.047	0.295

ϕ Calculated within a 300 meter buffer of the lakes.

Appendix E - Logistic Regression Models

#EWM code in R

ewmglmulti<-glmulti(ewi ~ StandElev + StandEwid + Standdi87 + StandAcc + StandGam + StandMtemp + StandMppt + StandRanp + StandDevO + StandDevM + StandDecF + StandMixF + StandSscr + StandOpen + StandAido + StandEndo + StandEnef + StandRpr + StandShei, data = EwmLogisticModel, level = 1, maxsize = 6, method = "h", crit=aicc, fitfunction = "glm", confetsize=200, report = TRUE, family = binomial)

#CLP code in R

clpglmulti<-glmulti(clp ~ StandPeri + StandElev + StandClpd + Standdi87 + StandMtemp + StandGam + StandDevo + StandDevM + StandDeci + StandSscr + StandShei + StandRpr + StandOpen, data = ClpLogisticModel, level = 1, maxsize = 6, method = "h", crit=aicc, fitfunction = "glm", confetsize=200, report = TRUE, family = binomial)

#Example Results - Top 200 Models for CLP Logistic Regression

1 clp ~ 1 + StandElev + StandGam + StandClpd 51.61387 0.029895523
2 clp ~ 1 + StandElev + StandGam + StandClpd + StandOpen 51.87369 0.026253473
3 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandOpen 52.02351 0.024358598
4 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd 52.55378 0.018685601
5 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO 52.85684 0.016058256
6 clp ~ 1 + StandElev + StandGam + StandClpd + StandSscr 52.97574 0.015131398
7 clp ~ 1 + StandElev + StandGam + StandClpd + StandSscr + StandOpen 53.00250 0.014930269
8 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevM 53.03522 0.014688024
9 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandSscr + StandOpen 53.22778 0.013339793
10 clp ~ 1 + StandElev + StandGam + StandClpd + StandShei 53.25941 0.013130494
11 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandOpen 53.38796 0.012313043
12 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandDevM + StandOpen 53.40438 0.012212392
13 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandDevM 53.40751 0.012193270
14 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevM + StandOpen 53.47051 0.011815222
15 clp ~ 1 + StandElev + StandGam + StandClpd + StandOpen + StandShei 53.63162 0.010900748
16 clp ~ 1 + StandElev + StandPeri + Standdi87 + StandGam + StandClpd + StandOpen 53.66084 0.010742634
17 clp ~ 1 + StandElev + StandGam + StandClpd + StandDecF 53.73499 0.010351648
18 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd 53.77711 0.010135938
19 clp ~ 1 + StandElev + StandGam + StandClpd + StandMtemp 53.78020 0.010120282
20 clp ~ 1 + StandElev + StandGam + StandClpd + StandRpr 53.78081 0.010117173
21 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevM + StandSscr 53.93269 0.009377329
22 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO + StandOpen 53.95276 0.009283730
23 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandDecF + StandOpen 53.98395 0.009140070
24 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevM + StandSscr + StandOpen 54.00902 0.009026206
25 clp ~ 1 + StandElev + StandGam + StandClpd + StandOpen + StandRpr 54.03472 0.008910943
26 clp ~ 1 + StandElev + StandGam + StandClpd + StandMtemp + StandOpen 54.06632 0.008771275
27 clp ~ 1 + StandElev + StandGam + StandClpd + StandDecF + StandOpen 54.07954 0.008713462
28 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandDevO 54.09108 0.008663366
29 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandOpen + StandShei 54.11759 0.008549280
30 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandDevO + StandOpen 54.14983 0.008412564
31 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO + StandSscr 54.16998 0.008328214
32 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandOpen + StandRpr 54.19939 0.008206677
33 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandDecF 54.26211 0.007953298
34 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandMtemp + StandOpen 54.26342 0.007948099
35 clp ~ 1 + StandElev + Standdi87 + StandGam 54.26487 0.007942337
36 clp ~ 1 + StandElev + StandGam + StandClpd + StandSscr + StandShei 54.39632 0.007437121
37 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandSscr 54.48017 0.007131760
38 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandShei 54.56748 0.006827124
39 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandDevM + StandOpen 54.67004 0.006485836
40 clp ~ 1 + StandElev + StandPeri + Standdi87 + StandGam + StandClpd 54.70291 0.006380112
41 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandMtemp 54.73620 0.006274784
42 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO + StandDevM 54.74680 0.006241633
43 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandRpr 54.74872 0.006235646
44 clp ~ 1 + StandElev + Standdi87 + StandGam + StandDevM 54.80454 0.006064023
45 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandShei 54.83055 0.005985670

46 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandDevO 54.85209 0.005921548
47 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevM + StandShei 54.86357 0.005887660
48 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO + StandRpr 54.87744 0.005846948
49 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandOpen + StandShei 54.88432 0.005826865
50 clp ~ 1 + StandElev + Standdi87 + StandGam + StandOpen 54.88941 0.005812075
51 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO + StandDecF 54.89788 0.005787497
52 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO + StandShei 54.97164 0.005577948
53 clp ~ 1 + StandElev + StandGam + StandClpd + StandSscr + StandOpen + StandRpr 54.98611 0.005537729
54 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevM + StandOpen + StandShei 54.99950 0.005500796
55 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandDevM + StandDecF 55.02659 0.005426776
56 clp ~ 1 + StandElev + StandGam + StandClpd + StandSscr + StandOpen + StandShei 55.03083 0.005415302
57 clp ~ 1 + StandElev + StandGam + StandClpd + StandDecF + StandSscr 55.03557 0.005402466
58 clp ~ 1 + StandElev + StandGam + StandClpd + StandMtemp + StandDevO 55.05028 0.005362870
59 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO + StandSscr + StandOpen 55.05379 0.005353473
60 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandSscr + StandOpen 55.06925 0.005312245
61 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandDevM + StandSscr 55.09244 0.005251026
62 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevM + StandRpr 55.09775 0.005237097
63 clp ~ 1 + StandElev + StandGam + StandClpd + StandSscr + StandRpr 55.13118 0.005150290
64 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandSscr 55.17672 0.005034330
65 clp ~ 1 + StandElev + StandGam + StandClpd + StandMtemp + StandSscr 55.17770 0.005031871
66 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandDevM 55.19279 0.004994044
67 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevM + StandDecF 55.19558 0.004987081
68 clp ~ 1 + StandElev + StandGam + StandClpd + StandMtemp + StandDevM 55.21198 0.004946359
69 clp ~ 1 + StandElev + StandGam + StandClpd + StandDecF + StandSscr + StandOpen 55.22427 0.004916061
70 clp ~ 1 + StandElev + StandGam + StandClpd + StandMtemp + StandSscr + StandOpen 55.23008 0.004901802
71 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandMtemp + StandDevM 55.32568 0.004673009
72 clp ~ 1 + StandElev + StandGam + StandSscr + StandOpen 55.38104 0.004545429
73 clp ~ 1 + StandElev + StandPeri + Standdi87 + StandGam + StandClpd + StandDevM 55.43461 0.004425295
74 clp ~ 1 + StandElev + StandGam + StandClpd + StandDecF + StandShei 55.43907 0.004415446
75 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandDecF + StandOpen 55.44955 0.004392367
76 clp ~ 1 + StandElev + StandGam + StandClpd + StandMtemp + StandShei 55.44998 0.004391404
77 clp ~ 1 + StandElev + StandGam + StandClpd + StandRpr + StandShei 55.45476 0.004380938
78 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandDevO + StandDevM 55.51501 0.004250931
79 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevM + StandSscr + StandShei 55.51753 0.004245565
80 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandDevO + StandOpen 55.53607 0.004206387
81 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandDevM + StandRpr 55.54057 0.004196947
82 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandOpen + StandRpr 55.57161 0.004132308
83 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandMtemp + StandOpen 55.62818 0.004017060
84 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandDevM + StandShei 55.63039 0.004012638
85 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO + StandOpen + StandShei 55.66564 0.003942524
86 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevM + StandOpen + StandRpr 55.68732 0.003900014
87 clp ~ 1 + StandElev + StandGam + StandClpd + StandMtemp + StandDevM + StandOpen 55.69563 0.003883846
88 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO + StandDevM + StandOpen 55.71158 0.003853000
89 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevM + StandDecF + StandOpen 55.71352 0.003849271
90 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandDevO + StandDecF 55.73261 0.003812693
91 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO + StandDevM + StandSscr 55.74815 0.003783189
92 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandSscr + StandShei 55.78053 0.003722429
93 clp ~ 1 + StandElev + StandGam + StandDevM + StandSscr + StandOpen 55.79196 0.003701216
94 clp ~ 1 + StandElev + Standdi87 + StandGam + StandDevM + StandOpen 55.81541 0.003658065
95 clp ~ 1 + StandElev + StandGam + StandClpd + StandOpen + StandRpr + StandShei 55.84347 0.003607108
96 clp ~ 1 + StandElev + StandGam + StandClpd + StandMtemp + StandOpen + StandShei 55.86492 0.003568628
97 clp ~ 1 + StandElev + StandGam + StandClpd + StandDecF + StandOpen + StandShei 55.86645 0.003565894
98 clp ~ 1 + StandElev + Standdi87 + StandGam + StandOpen + StandShei 55.89129 0.003521889
99 clp ~ 1 + StandElev + Standdi87 + StandGam + StandSscr 55.91305 0.003483777
100 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandDecF 55.92240 0.003467521
101 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandDevO + StandSscr 55.93046 0.003453584
102 clp ~ 1 + StandElev + StandGam + StandClpd + StandMtemp + StandDecF 55.93272 0.003449677
103 clp ~ 1 + StandElev + StandGam + StandClpd + StandDecF + StandRpr 55.93619 0.003443694
104 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandRpr 55.97568 0.003376364
105 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandMtemp 55.98082 0.003367700
106 clp ~ 1 + StandElev + StandGam + StandClpd + StandMtemp + StandRpr 55.98219 0.003365392
107 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevM + StandDecF + StandSscr 56.00713 0.003323687
108 clp ~ 1 + StandElev + Standdi87 + StandGam + StandDevM + StandSscr 56.02962 0.003286524
109 clp ~ 1 + StandElev + StandPeri + Standdi87 + StandGam + StandClpd + StandDevO 56.03500 0.003277696
110 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO + StandDecF + StandSscr 56.07594 0.003211290
111 clp ~ 1 + StandElev + Standdi87 + StandGam + StandDevM + StandOpen + StandShei 56.09650 0.003178442
112 clp ~ 1 + StandElev + StandGam + StandClpd + StandMtemp + StandDevM + StandSscr 56.13398 0.003119435
113 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevM + StandSscr + StandRpr 56.13968 0.003110560
114 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO + StandSscr + StandShei 56.14388 0.003104034
115 clp ~ 1 + StandElev + StandGam + StandOpen 56.14465 0.003102832

116 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandDevM + StandSscr 56.14774 0.003098042
117 clp ~ 1 + StandElev + Standdi87 + StandGam + StandDevO 56.17015 0.003063522
118 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO + StandDecF + StandOpen 56.18165 0.003045967
119 clp ~ 1 + StandElev + StandGam + StandClpd + StandMtemp + StandDevO + StandOpen 56.18458 0.003041499
120 clp ~ 1 + StandElev + Standdi87 + StandGam + StandDecF 56.19518 0.003025424
121 clp ~ 1 + StandElev + Standdi87 + StandGam + StandSscr + StandOpen 56.19562 0.003024758
122 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO + StandOpen + StandRpr 56.19602 0.003024152
123 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandDecF + StandSscr 56.19774 0.003021557
124 clp ~ 1 + StandElev + StandPeri + Standdi87 + StandGam + StandClpd + StandShei 56.24198 0.002955449
125 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandDevO + StandSscr 56.25976 0.002929290
126 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandDevO + StandRpr 56.26044 0.002928305
127 clp ~ 1 + StandElev + StandPeri + Standdi87 + StandGam + StandClpd + StandDecF 56.26275 0.002924912
128 clp ~ 1 + StandElev + StandGam + StandClpd + StandMtemp + StandOpen + StandRpr 56.27335 0.002909457
129 clp ~ 1 + StandElev + StandGam + StandClpd + StandDecF + StandOpen + StandRpr 56.27789 0.002902860
130 clp ~ 1 + StandElev + StandGam + StandClpd + StandMtemp + StandDecF + StandOpen 56.30905 0.002857990
131 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandMtemp + StandDevO 56.32539 0.002834723
132 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandDevO + StandShei 56.32923 0.002829286
133 clp ~ 1 + StandElev + Standdi87 + StandGam + StandShei 56.33649 0.002819038
134 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandSscr + StandShei 56.33900 0.002815506
135 clp ~ 1 + StandElev + StandPeri + Standdi87 + StandGam 56.35708 0.002790163
136 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO + StandSscr + StandRpr 56.36860 0.002774138
137 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandDevM + StandShei 56.39871 0.002732694
138 clp ~ 1 + StandElev + StandGam + StandClpd + StandMtemp + StandDevO + StandSscr 56.40916 0.002718445
139 clp ~ 1 + StandElev + StandGam + StandClpd + StandSscr + StandRpr + StandShei 56.41935 0.002704635
140 clp ~ 1 + StandElev + Standdi87 + StandGam + StandRpr 56.42205 0.002700981
141 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandDecF + StandShei 56.42211 0.002700909
142 clp ~ 1 + StandElev + Standdi87 + StandGam + StandMtemp 56.43419 0.002684637
143 clp ~ 1 + Standdi87 + StandGam + StandClpd + StandDevM 56.43888 0.002678345
144 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandMtemp + StandDecF 56.49650 0.002602284
145 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandDecF + StandRpr 56.50432 0.002592133
146 clp ~ 1 + StandElev + StandGam + StandSscr 56.51399 0.002579625
147 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandDevO + StandRpr 56.52080 0.002570865
148 clp ~ 1 + StandElev + StandGam + StandClpd + StandDecF + StandSscr + StandShei 56.53919 0.002547337
149 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandDevO + StandShei 56.54663 0.002537870
150 clp ~ 1 + StandElev + Standdi87 + StandGam + StandDevM + StandRpr 56.55432 0.002528133
151 clp ~ 1 + StandElev + Standdi87 + StandGam + StandDevM + StandDecF 56.56386 0.002516107
152 clp ~ 1 + StandPeri + Standdi87 + StandGam + StandClpd + StandOpen 56.56664 0.002512607
153 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO + StandDevM + StandRpr 56.57576 0.002501176
154 clp ~ 1 + StandElev + Standdi87 + StandGam + StandDevM + StandSscr + StandOpen 56.57929 0.002496761
155 clp ~ 1 + StandPeri + Standdi87 + StandGam + StandClpd + StandDevM + StandOpen 56.60854 0.002460522
156 clp ~ 1 + StandElev + StandGam + StandClpd + StandMtemp + StandSscr + StandShei 56.61344 0.002454498
157 clp ~ 1 + StandElev + StandGam 56.63923 0.002423044
158 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandSscr + StandRpr 56.65414 0.002405058
159 clp ~ 1 + StandElev + StandPeri + Standdi87 + StandGam + StandClpd + StandSscr 56.68050 0.002373564
160 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandMtemp + StandSscr 56.71049 0.002338236
161 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandDevO + StandDecF 56.71650 0.002331224
162 clp ~ 1 + Standdi87 + StandGam + StandClpd + StandDevO 56.72264 0.002324076
163 clp ~ 1 + StandElev + Standdi87 + StandGam + StandMtemp + StandDevM 56.74801 0.002294776
164 clp ~ 1 + StandElev + StandPeri + StandClpd + StandDevO 56.75623 0.002285368
165 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandDevO + StandDevM 56.77772 0.002260936
166 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandRpr + StandShei 56.78054 0.002257757
167 clp ~ 1 + StandElev + StandPeri + Standdi87 + StandGam + StandOpen 56.79464 0.002241890
168 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandMtemp + StandShei 56.80426 0.002231138
169 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO + StandDevM + StandDecF 56.85776 0.002172243
170 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO + StandDevM + StandShei 56.89231 0.002135040
171 clp ~ 1 + StandPeri + Standdi87 + StandClpd + StandDevO 56.90357 0.002123060
172 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO + StandDecF + StandRpr 56.90815 0.002118197
173 clp ~ 1 + StandElev + StandPeri + Standdi87 + StandGam + StandClpd + StandMtemp 56.90860 0.002117719
174 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandDecF + StandShei 56.91416 0.002111839
175 clp ~ 1 + StandElev + StandPeri + Standdi87 + StandGam + StandClpd + StandRpr 56.94528 0.002079233
176 clp ~ 1 + StandPeri + Standdi87 + StandGam + StandClpd + StandDevO 56.95562 0.002068513
177 clp ~ 1 + StandElev + Standdi87 + StandGam + StandClpd + StandMtemp + StandRpr 56.95859 0.002065446
178 clp ~ 1 + StandElev + StandGam + StandClpd + StandMtemp + StandDevO + StandDevM 56.98738 0.002035929
179 clp ~ 1 + StandElev + StandGam + StandSscr + StandShei 57.00498 0.002018094
180 clp ~ 1 + StandElev + Standdi87 + StandGam + StandDevM + StandShei 57.00680 0.002016252
181 clp ~ 1 + StandElev + Standdi87 + StandGam + StandDevO + StandDevM 57.00902 0.002014013
182 clp ~ 1 + StandElev + StandPeri + Standdi87 + StandGam + StandDevM 57.01014 0.002012889
183 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevM + StandRpr + StandShei 57.04222 0.001980864
184 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandRpr + StandShei 57.04551 0.001977603
185 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandMtemp + StandShei 57.06148 0.001961880

186 clp ~ 1 + StandElev + Standdi87 + StandGam + StandDecF + StandOpen 57.06196 0.001961403
187 clp ~ 1 + StandElev + Standdi87 + StandClpd + StandRpr 57.06588 0.001957564
188 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevM + StandDecF + StandShei 57.07804 0.001945701
189 clp ~ 1 + StandElev + Standdi87 + StandGam + StandMtemp + StandOpen 57.07992 0.001943871
190 clp ~ 1 + StandElev + StandGam + StandClpd + StandMtemp + StandDevO + StandRpr 57.08334 0.001940554
191 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandMtemp + StandDevO 57.08550 0.001938456
192 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO + StandDecF + StandShei 57.08760 0.001936420
193 clp ~ 1 + StandElev + Standdi87 + StandGam + StandOpen + StandRpr 57.09021 0.001933897
194 clp ~ 1 + StandElev + Standdi87 + StandGam + StandDevO + StandOpen 57.09469 0.001929565
195 clp ~ 1 + StandElev + StandGam + StandClpd + StandMtemp + StandDevM + StandShei 57.09574 0.001928559
196 clp ~ 1 + Standdi87 + StandGam + StandDevM 57.10003 0.001924423
197 clp ~ 1 + StandElev + StandGam + StandClpd + StandDevO + StandRpr + StandShei 57.10118 0.001923321
198 clp ~ 1 + StandElev + StandGam + StandClpd + StandMtemp + StandDevO + StandDecF 57.10835 0.001916434
199 clp ~ 1 + StandElev + StandClpd + StandSscr + StandRpr 57.15423 0.001872975
200 clp ~ 1 + StandElev + StandPeri + StandGam + StandClpd + StandDevM + StandRpr 57.15721 0.001870179

Appendix F – LR and GWLR Model Comparison Tables

Logistic and geographically weighted logistic regression model comparison for EWM presence-absence predictions.

Diagnostic Statistics	LR	GWLR	Difference
Model 1 Akaike's Information Criterion (AICc)	81.914	79.560	2.354
Deviance (-2 Log Likelihood)	66.952	57.526	9.426
R-square	0.586	0.597	-0.095
Percent correctly predicted (cut off at 0.5)	88.10%	89.68%	-1.587%
RMSE	0.168	0.148	0.020
Residual Global Moran's I	-0.012 (0.677)	-0.026 (0.465)	-0.014
Model 2 Akaike's Information Criterion (AICc)	90.397	84.240	5.106
Deviance (-2 Log Likelihood)	77.691	57.472	12.517
R-square	0.520	0.645	-0.251
Percent correctly predicted (cut off at 0.5)	86.508%	90.476%	-1.587%
RMSE	0.195	0.158	0.018
Residual Global Moran's I ‡	0.067 (0.004)	0.018 (0.313)	0.015
Model 3 Akaike's Information Criterion (AICc)	86.405	85.891	5.794
Deviance (-2 Log Likelihood)	71.456	66.552	13.820
R-square	0.442	0.589	-0.008
Percent correctly predicted (cut off at 0.5)	86.51%	87.30%	-3.175%
RMSE	0.179	0.170	0.029
Residual Global Moran's I	-0.011 (0.896)	-0.015 (0.778)	0.031
Model 4 Akaike's Information Criterion (AICc)	85.310	78.698	5.531
Deviance (-2 Log Likelihood)	72.604	49.505	12.484
R-square	0.551	0.694	-0.160
Percent correctly predicted (cut off at 0.5)	85.714%	89.68%	-2.381%
RMSE	0.183	0.132	0.024
Residual Global Moran's I	0.024 (0.216)	-0.050 (0.109)	0.000
Model 5 Akaike's Information Criterion (AICc)	83.398	82.648	0.750
Deviance (-2 Log Likelihood)	68.449	63.514	4.935
R-square	0.577	0.751	-0.174
Percent correctly predicted (cut off at 0.5)	88.89%	89.68%	-0.794%
RMSE	0.168	0.159	0.009
Residual Global Moran's I	-0.023 (0.559)	-0.025 (0.598)	-0.002

‡ z-score LR = 2.865 (< 1% chance clustered pattern is result of random chance).

Logistic and geographically weighted logistic regression model comparison for CLP presence-absence predictions.

	Diagnostic Statistics	LR	GWR	Difference
Model 1	Akaike's Information Criterion (AICc)	51.614	50.202	1.412
	Deviance (-2 Log Likelihood)	43.283	37.720	5.563
	R-square	0.530	0.590	-0.060
	Percent correctly predicted (cut off at 0.5)	94.44%	94.44%	0.000%
	RMSE	0.098	0.0949	0.003
	Residuals Global Moran's I	-0.010 (0.919)	-0.014 (0.790)	-0.004
Model 2	Akaike's Information Criterion (AICc)	53.210	53.117369	0.093
	Deviance (-2 Log Likelihood)	38.261	36.052524	2.208
	R-square	0.584	0.608	-0.024
	Percent correctly predicted (cut off at 0.5)	94.44%	94.44%	0.000%
	RMSE	0.090	0.0887	0.001
	Residuals Global Moran's I	-0.011 (0.913)	-0.011 (0.903)	0.000
Model 3	Akaike's Information Criterion (AICc)	54.265	55.616873	-1.352
	Deviance (-2 Log Likelihood)	45.934	31.136921	14.797
	R-square	0.501	0.662	-0.161
	Percent correctly predicted (cut off at 0.5)	92.06%	93.65%	-1.587%
	RMSE	0.104	0.0832	0.021
	Residuals Global Moran's I	0.000 (0.724)	-0.014 (0.855)	0.014
Model 4	Akaike's Information Criterion (AICc)	53.779	52.927282	0.852
	Deviance (-2 Log Likelihood)	43.280	37.737416	5.543
	R-square	0.530	0.590	-0.060
	Percent correctly predicted (cut off at 0.5)	94.44%	94.44%	0.000%
	RMSE	0.098	0.095	0.003
	Residuals Global Moran's I	-0.010 (0.913)	-0.012 (0.712)	0.002
Model 5	Akaike's Information Criterion (AICc)	57.575	54.178197	3.397
	Deviance (-2 Log Likelihood)	49.245	40.501423	8.743
	R-square	0.465	0.560	-0.095
	Percent correctly predicted (cut off at 0.5)	92.06%	94.44%	-2.38%
	RMSE	0.114	0.102	0.012
	Residuals Global Moran's I	0.011 (0.365)	-0.026 (0.481)	0.037

References

Adirondack Park Agency (APA). (2017). Adirondack Park Agency Maps and Geographic Information Systems. Retrieved from https://apa.ny.gov/gis/.

Adirondack Park Invasive Plant Program (APIPP). (2016). Retrieved from http://adkinvasives.com.

Aiken, S. G., Newroth, P. R., & Wile, I. (1979). THE BIOLOGY OF CANADIAN WEEDS.: 34. Myriophyllum spicatum L. *Canadian Journal of Plant Science*, *59*(1), 201-215.

Alberti, M., & Marzluff, J. M. (2004). Ecological resilience in urban ecosystems: linking urban patterns to human and ecological functions. *Urban ecosystems*, *7*(3), 241-265.

Babyak, M. A. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic medicine*, *66*(3), 411-421.

Barnes, M. A., Jerde, C. L., Keller, D., Chadderton, W. L., Howeth, J. G., & Lodge, D. M. (2013). Viability of aquatic plant fragments following desiccation. *Invasive Plant Science and Management*, *6*(2), 320-325.

Bruckerhoff, L., Havel, J., & Knight, S. (2015). Survival of invasive aquatic plants after air exposure and implications for dispersal by recreational boats. *Hydrobiologia, 746*(1), 113-121. doi:http://dx.doi.org/10.1007/s10750-014-1947-9

Buchan, L. A., & Padilla, D. K. (1999). Estimating the Probability of Long-Distance Overland Dispersal of Invading Aquatic Species. *Ecological applications*, *9*(1), 254-265.

Buchan, L. A., & Padilla, D. K. (2000). Predicting the likelihood of Eurasian watermilfoil presence in lakes, a macrophyte monitoring tool. *Ecological Applications*, *10*(5), 1442-1455.

Butchart, S. H., Walpole, M., Collen, B., Van Strien, A., Scharlemann, J. P., Almond, R. E., ... & Carpenter, K. E. (2010). Global biodiversity: indicators of recent declines. *Science*, *328*(5982), 1164-1168.

Burnham, K.P. and Anderson, D.R. (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. 2nd Edition. Springer-Verlag, New York.

Boylen, C. W., Eichler, L. W., & Madsen, J. D. (1999). Loss of native aquatic plant species in a community dominated by Eurasian watermilfoil. In *Biology, Ecology and Management of Aquatic Plants* (pp. 207-211). Springer Netherlands.

Boylen, C. W., Eichler, L. W., Bartkowski, J. S., & Shaver, S. M. (2006). Use of Geographic Information Systems to monitor and predict non-native aquatic plant dispersal through north-eastern North America. *Hydrobiologia, 570*(1), 243-248.

Cacho, O. J., Spring, D., Hester, S., & Mac Nally, R. (2010). Allocating surveillance effort in the management of invasive species: a spatially-explicit model. *Environmental Modelling & Software, 25*(4), 444-454.

Calcagno, Vincent. (2015). Package 'glmulti'. Retrieved from https://cran.r-project.org/web/packages/glmulti/glmulti.pdf

Calcagno, V. and Mazancourt, C. (2010). glmulti: An R Package for Easy Automated Model Selection with Generalized Linear models. Journal of Statistical Software, 34(12).

Capers, R. S., Selsky, R., Bugbee, G. J., & White, J. C. (2007). AQUATIC PLANT COMMUNITY INVASIBILITY AND SCALE-DEPENDENT PATTERNS IN NATIVE AND INVASIVE SPECIES RICHNESS. *Ecology, 88*(12), 3135-3143.

Carpenter, S. R., and K. L. Cottingham. 1997. Resilience and restoration of lakes. Conservation Ecology [online]**1**(1): 2. Available from the Internet. URL: http://www.consecol.org/vol1/iss1/art2/

Catling, P. M., & Dobson, I. (1985). THE BIOLOGY OF CANADIAN WEEDS.: 69. Potamogeton crispus L. *Canadian journal of plant science*, *65*(3), 655-668.

Charlton, M. and Fotheringham, A.S. n.d. Geographically weighted regression: A tutorial on using GWR in ArcGIS 9.3. Retrieved from http://www.geos.ed.ac.uk/~gisteac/fspat/gwr/gwr_arcgis/GWR_Tutorial.pdf

Cheruvelil, K. S., & Soranno, P. A. (2008). Relationships between lake macrophyte cover and lake and landscape features. *Aquatic Botany, 88*(3), 219-227.

Couch, Richard, and E. Nelson. "Myriophyllum spicatum in North America."*Proceedings of the First International Symposium on watermilfoil*. 1985.

Crall, A. W., Newman, G. J., Stohlgren, T. J., Holfelder, K. A., Graham, J., & Waller, D. M. (2011). Assessing citizen science data quality: an invasive species case study. *Conservation Letters*, *4*(6), 433-442.

Darbyson, E., Locke, A., Hanson, J. M., & Willison, J. M. (2009). Marine boating habits and the potential for spread of invasive species in the Gulf of St. Lawrence. *Aquatic Invasions*, *4*(1), 87-94.

Densham, P. J. (1991). Spatial decision support systems. *Geographical information systems: Principles and applications, 1*, 403-412.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ... & Münkemüller, T. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27-46.

Drake, J. M., & Lodge, D. M. (2004). Global hot spots of biological invasions: evaluating options for ballast–water management. *Proceedings of the Royal Society of London B: Biological Sciences*, *271*(1539), 575-580.

Dunbar, G. (2009). Management plan for eurasian watermilfoil (Myriophyllum spicatum) in the Okanagan, British Columbia. *Okanagan Basin Water Board*.

Erickson, B., & Nosanchuk, T. (1992). *Understanding data*. McGraw-Hill Education (UK).

ESRI. (2017). How Spatial Autocorrelation (Global Moran's I) works. Retrieved from http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/h-how-spatial-autocorrelation-moran-s-i-spatial-st.htm

Evans, C. A., Kelting, D. L., Forrest, K. M., & Steblen, L. E. (2011). Fragment viability and rootlet formation in Eurasian watermilfoil after desiccation. *Journal of Aquatic Plant Management*, *48*, 57-62.

Field, A. (2009). *Discovering statistics using SPSS*. Sage publications.

Fielding, A. H., & Haworth, P. F. (1995). Testing the generality of bird-habitat models. *Conservation biology*, *9*(6), 1466-1481.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons.

Fotheringham, S., Kelly, M., and Charlton, M. (n.d.) Model Selection in Geographically Weighted Regression. Retrieved from http://www.isprs.org/proceedings/xxxviii/part2/presentations/s10/fotheringham.pdf

Goodchild, M. F. (2003). Geographic information science and systems for environmental management. *Annual Review of Environment and Resources*, *28*.

Green, A. J. (2016). The importance of waterbirds as an overlooked pathway of invasion for alien species. *Diversity and Distributions*, *22*(2), 239-247.

Guerry, A. D., & Hunter, M. L. (2002). Amphibian distributions in a landscape of forests and agriculture: an examination of landscape composition and configuration. *Conservation Biology*, *16*(3), 745-754.

Gumpertz, L., & Pye, M. (2000). Logistic regression for southern pine beetle outbreaks with spatial and temporal autocorrelation. *Forest Science*, *46*(1), 95-107.

Hair Jr., J., Black W.C., Babin B.J., Anderson R.E. (2010). Multivariate Data Analysis. Pearson Prentice Hall.

Hastings, A., Cuddington, K., Davies, K. F., Dugaw, C. J., Elmendorf, S., Freestone, A., ... & Melbourne, B. A. (2005). The spatial spread of invasions: new developments in theory and evidence. *Ecology Letters*, *8*(1), 91-101.

Heiskary, S., & Valley, R. D. (2012). Curly-leaf pondweed trends and interrelationships with water quality. *Transition*, *525*(10,580), 5-0.

Herborg, L. M., O'Hara, P., & Therriault, T. W. (2009). Forecasting the potential distribution of the invasive tunicate Didemnum vexillum. *Journal of Applied Ecology*, *46*(1), 64-72.

Higgins, S. I., & Richardson, D. M. (1996). A review of models of alien plant spread. *Ecological modelling*, *87*(1-3), 249-265.

Hopkins, K. D., & Weeks, D. L. (1990). Tests for normality and measures of skewness and kurtosis: Their place in research reporting. *Educational and Psychological Measurement*, *50*(4), 717-729.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2000). *Applied logistic regression*. John Wiley & Sons.

Hulme, P. E. (2009). Trade, transport and trouble: managing invasive species pathways in an era of globalization. *Journal of applied ecology*, *46*(1), 10-18.

Jenkins, J. (2004). *The Adirondack Atlas: A Geographic Portrait of the Adirondack Park.* Syracuse University Press.

Jenkins, P. 2001. Economic Impacts of Aquatic Nuisance Species in the Great Lakes. A report prepared by Philip Jenkins and Associates, Ltd. for Environment Canada. Burlington, Ontario.

Jerde, C. L., Barnes, M. A., DeBuysser, E. K., Noveroske, A., Chadderton, W. L., & Lodge, D. M. (2012). Eurasian watermilfoil fitness loss and invasion potential following desiccation during simulated overland transport. *Aquatic Invasions*, *7*(2).

Jiménez-Valverde, A., Peterson, A. T., Soberón, J., Overton, J. M., Aragón, P., & Lobo, J. M. (2011). Use of niche models in invasive species risk assessments. *Biological invasions*, *13*(12), 2785-2797.

Johnson, L. E., A. Ricciardi, and J. T. Carlton. (2001). Overland dispersal of aquatic invasive species: a risk assessment of transient recreational boating. *Ecological Applications 11*:1789-1799.

Johnson, R. L., Van Dusen, P. J., Toner, J. A., & Hairston, N. G. (2000). Eurasian watermilfoil biomass associated with insect herbivores in New York. *Journal of Aquatic Plant Management*, *38*, 82-88.

Keller, R.P, and Lodge, D.M. (2007). Species Invasions from Commerce in Live Aquatic Organisms: Problems and Possible Solutions. *BioScience, 57 (5),* 428-436.

Kilian, J. V., Klauda, R. J., Widman, S., Kashiwagi, M., Bourquin, R., Weglein, S., & Schuster, J. (2012). An assessment of a bait industry and angler behavior as a vector of invasive species. *Biological Invasions*, *14*(7), 1469-1481.

Kratz, T., Webster, K., Bowser, C., Maguson, J., and Benson B. (1997). The influence of landscape position on lakes in northern Wisconsin.*Freshwater Biology*, *37*(1), 209-217.

Krecker, F.H. (1939). A comparative study of the animal population of certain submerged aquatic plants. *Ecoogy*, *20* (4), 553 – 562.

LCLT (Lake Champlain Land Trust). (2017). Lake Champlain Facts. Retrieved from http://www.lclt.org/about-lake-champlain/lake-champlain-facts/

Les, D. H., & Mehrhoff, L. J. (1999). Introduction of nonindigenous aquatic vascular plants in southern New England: a historical perspective. *Biological Invasions*, *1*(2-3), 281-300.

Leung, B., Bossenbroek, J. M., & Lodge, D. M. (2006). Boats, pathways, and aquatic biological invasions: estimating dispersal potential with gravity models. *Biological Invasions*, *8*(2), 241-254.

Lillie, R. A., & Barko, J. W. (1990). Influence of Sediment and Groundwater on the Distribution and Biomass of Myriophyllum spicatum L. in Devil's Lake, Wisconsin. *Journal of Freshwater Ecology*, *5*(4), 417-426.

Lovell, S. J., Stone, S. F., & Fernandez, L. (2006). The economic impacts of aquatic invasive species: a review of the literature. *Agricultural and Resource Economics Review*, *35*(1), 195.

Lowe S., Browne M., Boudjelas S., De Poorter M. (2000) 100 of the World's Worst Invasive Alien Species A selection from the Global Invasive Species Database. Published by The Invasive Species Specialist Group (ISSG) a specialist group of the Species Survival Commission (SSC) of the World Conservation Union (IUCN), 12pp. First published as special lift-out in Aliens 12, December 2000. Updated and reprinted version: November 2004.

Lee, S. B., Hwang, H. S., & Sung, H. C. (2009). Landscape ecological approach to the relationships of land use patterns in watersheds to water quality characteristics. *Landscape and Urban Planning*, *92*(2), 80-89.

Mack, R. N., Simberloff, D., Mark Lonsdale, W., Evans, H., Clout, M., & Bazzaz, F. A. (2000). Biotic invasions: causes, epidemiology, global consequences, and control. *Ecological applications*, *10*(3), 689-710.

Marquaridt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, *12*(3), 591-612.

Mataraza, L. K., J. B. Terrell, A. B. Munson and D. E. Canfield, Jr. (1999). Changes in submersed macrophytes in relation to tidal storm surges. *J. Aquat. Plant Manage. 37*: 3-12.

Maynooth University. (2017). Geographically weighted modelling. Retrieved from http://gwr.maynoothuniversity.ie/

Madsen, J. D. (1998). Predicting invasion success of Eurasian watermilfoil. *Journal of Aquatic Plant Management*, *36*(2832), 122134.

Madsen, J. D., Eichler, L. W., & Boylen, C. W. (1988). Vegetative spread of Eurasian watermilfoil in Lake George, New York. *Journal of Aquatic Plant Management*, *26*(2), 47-50.

Madsen, J. D., Sutherland, J. W., Bloomfield, J. A., Eichler, L. W., & Boylen, C. W. (1991). The decline of native vegetation under dense Eurasian watermilfoil canopies. *Journal of Aquatic Plant Management*, *29*, 94-99.

Matthews, S. A., & Yang, T. C. (2012). Mapping the results of local statistics: Using geographically weighted regression. *Demographic research*, *26*, 151.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior.

McGarigal, K. (2009). Lecture on *Landscape Metrics for Categorical Map Patterns* [PDF]. Personal Collection of K. McGarigal, University of Massachusetts Amherst, Amherst, MA. Retrieved from: http://www.umass.edu/landeco/teaching/landscape_ecology/schedule/chapter9_metrics.pdf

McGarigal, K., & Marks, B. J. (1995). Spatial pattern analysis program for quantifying landscape structure. *Gen. Tech. Rep. PNW-GTR-351. US Department of Agriculture, Forest Service, Pacific Northwest Research Station.*

Miller, James H. 2003**.** Nonnative invasive plants of southern forests: a field guide for identification and control. Gen. Tech. Rep. SRS-62. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 93p.

Menard, Scott. (2000). Coefficients of determination for multiple logistic regression analysis. American Statistician, 54(1), 17-24.

Moody, M. L., Palomino, N., Weyl, P. S., Coetzee, J. A., Newman, R. M., Harms, N. E., ... & Thum, R. A. (2016). Unraveling the biogeographic origins of the Eurasian watermilfoil (Myriophyllum spicatum) invasion in North America. *American journal of botany*, *103*(4), 709-718.

Nichols, S. A., & Shaw, B. H. (1986). Ecological life histories of the three aquatic nuisance plants, Myriophyllum spicatum, Potamogeton crispus and Elodea canadensis. *Hydrobiologia*, *131*(1), 3-21.

National Center for Education Statistics (NCES). 2017. Learn by Doing: Running a logistic regression and interpreting results. Retrieved from https://nces.ed.gov/datalab/powerstats/tutorials/PS_creating_logistic_regression.pdf

NISC (National Invasive Species Council). (2006). Invasive Species Definition Clarification and Guidance White Paper. Retrieved from https://www.invasivespeciesinfo.gov/docs/council/isacdef.pdf

NISC (National Invasive Species Council). (2017). Executive Orders. Retrieved from https://www.invasivespeciesinfo.gov/laws/execorder.shtml#eo13112

NISIC (National Invasive Species Information Center). (2017). Agencies and Organizations. Retrieved from https://www.invasivespeciesinfo.gov/resources/orgfed.shtml

Neubert, M. G., & Parker, I. M. (2004). Projecting rates of spread for invasive species. *Risk Analysis*, *24*(4), 817-831.

Nichols, S. A., & Shaw, B. H. (1986). Ecological life histories of the three aquatic nuisance plants, Myriophyllum spicatum, Potamogeton crispus and Elodea canadensis. *Hydrobiologia*, *131*(1), 3-21.

NY DEC (New York Department of Environmental Conservation). (2017). Fish Stocking Lists. Retrieved from http://www.dec.ny.gov/outdoor/30467.html.

O'Reilly, Neal; Ehlinger, Timothy; and Shaker, Richard, "The Development and Evaluation of Methods for Quantifying Risk to Fish in Warm-water Streams of Wisconsin Using Self-Organized Maps: Influences of Watershed and Habitat Stressors" (2007). Center for Urban Environmental Studies Publications. Paper 14.

O'Sullivan, D., & Unwin, D. J. (2010). Geographic Information Analysis and Spatial Data. *Geographic Information Analysis, Second Edition*.

Office of Technology Assessment. U.S. Congress (OTA). 1993. Harmful NonIndigenous Species in the United States. OTA Publication OTA-F-565. US Government Printing Office, Washington DC: Availability: http://www.wws.princeton.edu:80/~ota/disk1/1993/9325_n.html

Orloff, J. & Bloom, J. *18.05 Introduction to Probability and Statistics.*Spring 2014. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu. License: Creative Commons BY-NC-SA.

Oxford Economics. (2015). The Economic Impact of Tourism in New York. Retrieved from http://www.roostadk.com/wp-content/uploads/2016/08/NYS-Tourism-Impact-Adirondacks-2015.pdf

Peng, C. Y. J., & So, T. S. H. (2002). Logistic regression analysis and reporting: A primer. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, *1*(1), 31-70.

Pimentel, D., Zuniga, R., & Morrison, D. (2005). Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecological economics*, *52*(3), 273-288.

Prasad, A. M., Iverson, L. R., Peters, M. P., Bossenbroek, J. M., Matthews, S. N., Sydnor, T. D., & Schwartz, M. W. (2010). Modeling the invasive emerald ash borer risk of spread using a spatially explicit cellular model. *Landscape ecology*, *25*(3), 353-369.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, *2*(1), 21-33.

Robinson, M. (2002). Eurasian Watermilfoil: An Invasive Aquatic Plant. Retrieved from http://www.mass.gov/eea/docs/dcr/watersupply/lakepond/factsheet/eurasian-milfoil.pdf

Rothlisberger, J. D., Chadderton, W. L., McNulty, J., & Lodge, D. M. (2010). Aquatic invasive species transport via trailered boats: what is being moved, who is moving it, and what can be done. *Fisheries*, *35*(3), 121-132.

Rupert, M.G., Cannon, S.H., Gartner, J.E., Michael, J.A., and Helsel, D.R., 2008, Using logistic regression to predict the probability of debris flows in areas burned by wildfires, southern California, 2003–2006: U.S. Geological Survey Open-File Report 2008–1370, 9 p.

Rockwell, H.W. Jr. 2003. "Summary of a Survey of the Literature on the Economic Impact of Aquatic Weeds." Report for the Aquatic Ecosystem Restoration Foundation. August. http://www.aquatics.org/pubs/economics.htm.

Rose, K. C., Greb, S. R., Diebel, M., & Turner, M. G. (2016). Annual precipitation regulates spatial and temporal drivers of lake water clarity. *Ecological Applications*.

Shaker, R.R., & Rapp, C.J. (2013). Investigating Aquatic Invasive Species Propagation within the Adirondack Region of New York: A Lake and Landscape Approach. *Papers in Applied Geography,* 36, 200-209.

Shaker, R. R., Rapp, C. J., & Yakubov, A. D. (2013). Examining patterns of aquatic invasion within the Adirondacks: an OLS and GLM approach. *Middle States Geographer*, *46*, 1-11.

Shaker, R.R. (2015) The well-being of nations: an empirical assessment of sustainable urbanization for Europe. *International Journal of Sustainable Development & World Ecology, 22 (5),* 375-387.

Shaker, R. R., Yakubov, A. D., Nick, S. M., Vennie-Vollrath, E., Ehlinger, T. J., & Wayne Forsythe, K. (2017). Predicting aquatic invasion in Adirondack lakes: a spatial analysis of lake and landscape characteristics. *Ecosphere*,*8*(3).

Shalizi, Cosma R. (2017). Advanced Data Analysis from an Elementary Point of View. Cambridge University Press.

Smith, C.S., and Barko, J.W. (1990). Ecology of Eurasian watermilfoil. *Journal of Aquatic Plant Management, 28*:55 – 64.

Stoltzfus, J. C. (2011). Logistic regression: a brief primer. *Academic Emergency Medicine*, *18*(10), 1099-1104.

Stuckey R. L. (1979). Distributional history of Potamogeton crispus (Curly pondweed) in North America. *Bartonia 46*: 22–42

Symonds, M.R.E. and Moussalli, A. (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. Behav Ecol Sociobiol, 65, 13-21.

Tobler, W. R. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46: 234–40.

Tracy, M., Montante, J. M., Allenson, T. E., & Hough, R. A. (2003). Long-term responses of aquatic macrophyte diversity and community structure to variation in nitrogen loading. *Aquatic Botany*, *77*(1), 43-52.

USDA (United States Department of Agriculture). (2017). GeoSpatialDataGateway. Retrieved from https://datagateway.nrcs.usda.gov/GDGOrder.aspx?order=QuickState

USGS (U.S. Geological Survey). 2014. The national map. Earth Resources Observation and Science (EROS) Center. Sioux Falls, San Diego, USA. https://nationalmap.gov/

USGS (United States Geological Survey). (2017). Nonindigenous Aquatic Species. Retrieved from https://nas.er.usgs.gov/graphs/All.aspx

Walsh, S. J., McCleary, A. L., Mena, C. F., Shao, Y., Tuttle, J. P., González, A., & Atkinson, R. (2008). QuickBird and Hyperion data analysis of an invasive plant species in the Galapagos Islands of Ecuador: Implications for control and land use management. *Remote Sensing of Environment*, *112*(5), 1927-1941.

Williams, G., Layman, K. L., & Stefan, H. G. (2004). Dependence of lake ice covers on climatic, geographic and bathymetric variables. *Cold Regions Science and Technology*, *40*(3), 145-164.

With, K. A. (2002). The landscape ecology of invasive spread. *Conservation Biology*, *16*(5), 1192-1203.

Vander Zanden, M. J., & Olden, J. D. (2008). A management framework for preventing the secondary spread of aquatic invasive species. *Canadian Journal of Fisheries and Aquatic Sciences*, *65*(7), 1512-1522.

Zhu, B., & Georgian, S. E. (2014). Interactions between invasive Eurasian watermilfoil and native water stargrass in Cayuga Lake, NY, USA. *Journal of Plant Ecology*, rtt063.

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*(1), 3-14.

References of studies using GWLR

Atkinson, P. M., German, S. E., Sear, D. A., & Clark, M. J. (2003). Exploring the relations between riverbank erosion and geomorphological controls using geographically weighted logistic regression. *Geographical Analysis*, *35*(1), 58-82.

Carrel, M., Escamilla, V., Messina, J., Giebultowicz, S., Winston, J., Yunus, M., ... & Emch, M. (2011). Diarrheal disease risk in rural Bangladesh decreases as tubewell density increases: a zero-inflated and geographically weighted analysis. *International journal of health geographics*, *10*(1), 41.

Chalkias, C., Kalogirou, S., & Ferentinou, M. (2014). Landslide susceptibility, Peloponnese peninsula in south Greece. *Journal of Maps*, *10*(2), 211-222.

Comber, A., Fisher, P., Brunsdon, C., & Khmag, A. (2012). Spatial analysis of remote sensing image classification accuracy. *Remote Sensing of Environment*, *127*, 237-246.

Feuillet, T., Coquin, J., Mercier, D., Cossart, E., Decaulne, A., Jónsson, H. P., & Sæmundsson, Þ. (2014). Focusing on the spatial non-stationarity of landslide predisposing factors in northern Iceland: Do paraglacial factors vary over space?. *Progress in Physical Geography*, *38*(3), 354-377.

Goovaerts, P., Xiao, H., Adunlin, G., Ali, A., Tan, F., Gwede, C. K., & Huang, Y. (2015). Geographically-weighted regression analysis of percentage of late-stage prostate cancer diagnosis in Florida. *Applied Geography*, *62*, 191-200.

Goovaerts, P., Wobus, C., Jones, R., & Rissing, M. (2016). Geospatial estimation of the impact of Deepwater Horizon oil spill on plant oiling along the Louisiana shorelines. *Journal of environmental management*, *180*, 264-271.

Guo, F., Selvalakshmi, S., Lin, F., Wang, G., Wang, W., Su, Z., & Liu, A. (2016). Geospatial information on geographical and human factors improved anthropogenic fire occurrence modeling in the Chinese boreal forest. *Canadian Journal of Forest Research*, *46*(4), 582-594.

Han, H., Jang, K. M., & Chung, J. S. (2017). Selecting suitable sites for mountain ginseng (Panax ginseng) cultivation by using geographically weighted logistic regression. *Journal of Mountain Science*, *14*(3), 492-500.

Li, H., Wei, Y. H. D., & Huang, Z. (2014). Urban land expansion and spatial dynamics in globalizing shanghai. *Sustainability*, *6*(12), 8856-8875.

Likongwe, P., Kihoro, J., Ngigi, M., & Jamu, D. (2015). Modeling spatial non-stationarity of Chambo in South-East Arm of Lake Malawi. *Asian Journal of Applied Science and Engineering*, *4*(2), 81-90.

Liu, Z., & Robinson, G. M. (2016). Residential development in the peri-urban fringe: The example of Adelaide, South Australia. *Land Use Policy*, *57*, 179-192.

Luo, J., & Wei, Y. D. (2009). Modeling spatial variations of urban growth patterns in Chinese cities: the case of Nanjing. *Landscape and Urban Planning*, *91*(2), 51-64.

Koutsias N., Martinez, J., Chuvieco, E., and Allgower, B. (2005). Modeling wildland fire occurrence in Southern Europe by a Geographically Weighted Regression Approach, eds J.D. Riva, F. Perez-Cabello and E. Chuevieco, Proceedings of the 5th International Workshop on Remote Sensing and GIS Applications to Forest Fire Management: Fire Effects Assessment. Pp: 57 - 60.

Mcnew, L. B., Gregory, A. J., & Sandercock, B. K. (2013). Spatial heterogeneity in habitat selection: Nest site selection by greater prairie-chickens. *The Journal of Wildlife Management*, *77*(4), 791-801.

Saefuddin, A., Setiabudi, N. A., & Fitrianto, A. (2012). On comparison between logistic regression and geographically weighted logistic regression: with application to Indonesian poverty data. *World Applied Sciences Journal*,*19*(2), 205-210.

Rodrigues, M., de la Riva, J., & Fotheringham, S. (2014). Modeling the spatial variation of the explanatory factors of human-caused wildfires in Spain using geographically weighted logistic regression. *Applied Geography*, *48*, 52-63.

Schultz, C., Alegría, A. C., Cornelis, J., & Sahli, H. (2016). Comparison of spatial and aspatial logistic regression models for landmine risk mapping. *Applied Geography*, *66*, 52-63.

Shafizadeh-Moghadam, H., & Helbich, M. (2015). Spatiotemporal variability of urban growth factors: A global and local perspective on the megacity of Mumbai. *International Journal of Applied Earth Observation and Geoinformation*, *35*, 187-198.

Stohlgren, T. J., & Schnase, J. L. (2006). Risk analysis for biological hazards: what we need to know about invasive species. *Risk analysis*, *26*(1), 163-173.

Wimberly, M. C., Yabsley, M. J., Baer, A. D., Dugan, V. G., & Davidson, W. R. (2008). Spatial heterogeneity of climate and land-cover constraints on distributions of tick-borne pathogens. *Global Ecology and Biogeography*, *17*(2), 189-202.

Windle, M.J.S., Rose G.A., Devillers R., and Fortin M., (2009). Exploring spatial non-stationarity of fisheries survey data using geographically weighted regression (GWR): an example from the Northwest Atlantic.

Wu, L., Deng, F., Xie, Z., Hu, S., Shen, S., Shi, J., & Liu, D. (2016). Spatial Analysis of Severe Fever with Thrombocytopenia Syndrome Virus in China Using a Geographically Weighted Logistic Regression Model. *International Journal of Environmental Research and Public Health*, *13*(11), 1125.

Wu, W., & Zhang, L. (2013). Comparison of spatial and non-spatial logistic regression models for modeling the occurrence of cloud cover in north-eastern Puerto Rico. *Applied Geography*, *37*, 52-62.

Valley, R. D., & Heiskary, S. (2012). Short-term declines in curlyleaf pondweed in Minnesota: potential influences of snowfall. *Lake and reservoir management*, *28*(4), 338-345.

Zhang, L., Wei, Y. D., & Meng, R. (2017). Spatiotemporal Dynamics and Spatial Determinants of Urban Growth in Suzhou, China. *Sustainability*, *9*(3), 393.

Zhang, M., Cao, X., Peng, L., & Niu, R. (2016). Landslide susceptibility mapping based on global and local logistic regression models in Three Gorges Reservoir area, China. *Environmental Earth Sciences*, *75*(11), 1-11.

Zhou, Y. B., Wang, Q. X., Liang, S., Gong, Y. H., Yang, M. X., Chen, Y., ... & Yang, Y. (2015). Geographical variations in risk factors associated with HIV infection among drug users in a prefecture in Southwest China. *Infectious diseases of poverty*, *4*(1), 38.

Yoneoka, D., Saito, E., & Nakaoka, S. (2016). New algorithm for constructing area-based index with geographical heterogeneities and variable selection: An application to gastric cancer screening. *Scientific reports*, *6*.